A Numerical Analysis of the Effect of Sampling of Alternatives in Discrete Choice Models

Sriharsha Nerella and Chandra R. Bhat
The University of Texas at Austin, Department of Civil Engineering
1 University Station C1761, Austin, Texas, 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
E-mail: hi_harsha@yahoo.com, bhat@mail.utexas.edu

TRB 2004: FOR PRESENTATION AND PUBLICATION
TRB Paper # 04-3196
Final Submission Date: March 30, 2004
Word Count: 7,690

ABSTRACT
A large number of alternatives characterize the choice set in many activity and travel choice contexts. Analysts generally sample alternatives from the choice set in such situations because estimating models from the full choice set can be very expensive or even prohibitive. This paper undertakes numerical experiments to examine the effect of the sample size of alternatives on model performance for both an MNL model (for which consistency with a subset of alternatives is guaranteed) and a mixed multinomial logit model (for which no consistency result holds).

1. INTRODUCTION

Several of the activity and travel decisions made by individuals, such as travel mode choice, activity participation location choice, residential location choice, and route choice, are discrete in nature. This recognition has led to the widespread use of discrete choice models in travel demand modeling. Almost all of these discrete choice models are based on the Random Utility Maximization (RUM) hypothesis, which assumes that a decision-making agent's choice is a reflection of underlying preferences for each of the available alternatives, and that the agent selects the alternative with the highest preference or utility. The underlying preferences are random to the analyst, because s/he does not observe all the factors considered by the decision-maker in the choice process.

An issue that arises in the RUM-based discrete choice modeling of many activity and travel related dimensions is the large number of alternatives in the choice set. For example, in an activity participation location or residential choice situation, a decision-maker can potentially have anywhere between a few hundreds of choice alternatives (if an aggregate spatial unit such as neighborhoods or traffic analysis zones is used to characterize the alternatives) to hundreds of thousands of choice alternatives (if a fine spatial resolution such as land parcels is used to characterize the alternatives). Similarly, in a route choice decision context, a traveler potentially has an infinite number of routes to choose from to travel to his/her desired location for activity participation. In such large choice set situations, it is challenging to consider all the alternatives during estimation because of the substantial effort that would be entailed in assembling the relevant dataset. The computational burden can also be an important consideration in estimation with a very large set of alternatives.[1]

The challenge of estimating choice models with a huge set of alternatives has led researchers to explore and apply methods to enable consistent estimation with only a subset of alternatives (see Table 1 for a list of studies that have used a subset of alternatives rather than the complete choice set). McFadden (*3*) proved that, in the case of the multinomial logit model (MNL), it is straightforward to consistently estimate parameters from a sample of alternatives by maximizing a conditional likelihood function which also has an MNL form. This is a neat theoretical result and is associated with the independence from irrelevant alternatives (IIA) property of the multinomial logit model. However, there has been no systematic numerical analysis, to our knowledge, examining how the sample size of alternatives affects the empirical accuracy and efficiency of the estimated parameters.

Another issue in choice situations with a large number of alternatives is the case when non-MNL models are used. The MNL model, while simple and elegant in structure, is saddled with the IIA property, which can be behaviorally unrealistic in many choice situations. For example, in an activity participation location or residential choice situation, it is possible (if not very likely) that the utility of spatial alternatives close to each other will have a higher degree of sensitivity due to common unobserved spatial elements. A common specification in the spatial

[1] On a philosophical note, one could argue that individuals are limited information processors, and do not consider more than a few alternatives in any choice situation. Clearly, this is an important research issue within the broad area of understanding the choice set generation process. However, in the absence of a clear understanding of the choice set generation process, the most common practical assumption is that all alternatives in the universal choice set are available. The current paper is positioned within the framework of this assumption of full choice set consideration. However, it should be emphasized that the paper provides guidance even for modeling frameworks that incorporate choice set formation explicitly [for example, see Swait (*1*); Basar and Bhat (*2*)]. In such frameworks, the model takes a non-MNL form with the full choice set representing the universal choice set from which some alternatives are considered by the decision maker. To the extent that the universal choice set is very large, the analyst may want to reduce the universal choice set size to something manageable for each decision maker. This situation is mimicked by our analysis in this paper with non-MNL models.

analysis literature for capturing such spatial correlation is to allow contiguous alternatives to be correlated (*4*). Similarly, in a route choice context, routes with overlapping links are likely to have a higher sensitivity between each other compared to paths with little or no overlap. A common specification, therefore, in route choice models is to assume that the covariance of path utilities is proportional to the overlap length (*5*). In these and other choice situations, the use of the MNL model is clearly not appropriate, though the analytic elegance and ability to sample alternatives within the MNL framework has led to its continued use in the literature. Recent simulation-related and GEV-based model developments, however, are very rapidly liberating the analyst from using restrictive model forms such as the MNL. But, theoretically speaking, sampling of alternatives does not provide consistent parameter estimates in these more advanced model forms. Thus, the dilemma for the analyst is whether to impose the unrealistic MNL structure at the outset or use a more realistic structure and then potentially "undo" the advantage of the richer structure by sampling of alternatives.

The discussion above provides the motivation for the current research. Specifically, this paper has two objectives. The first objective is to examine the effect of the sampling size of alternatives on the empirical accuracy and efficiency of estimated parameters (and other relevant fit statistics) in the context of the MNL model. While McFadden's (*3*) result shows theoretically that any sample size of alternatives will provide consistent estimates in the MNL framework, the question of how many alternatives to select is still an empirical one. The second objective is to assess the impact of the sampling size of alternatives on the empirical accuracy and efficiency of parameter and fit statistics in the context of non-MNL models. In such models, it is theoretically known that sampling of alternatives does not work, but the question is: Is there a certain size of alternatives that makes the results from the sample of alternatives close enough (empirically speaking) to the true values obtained from the full choice set?

A few notes are in order before we proceed. First, we use the mixed multinomial logit (MMNL) form as the representative structure for the non-MNL forms in this paper. This is because the MMNL model is a very flexible discrete choice structure, is easy to estimate, and is becoming the method of preference for accommodating behaviorally realistic structures. Second, our assessment of the effect of sample size of alternatives on model performance is based on numerical experiments. Third, the results from this paper should be viewed as providing guidance to the analyst when confronted with a choice situation with a large number of alternatives. The results should not be viewed as "absolute rules" since each empirical context is likely to be unique and different from others. It is simply impossible in a numerical experiment to consider all the situations that may arise in reality, including combinations of different sample sizes of observations, different numbers of alternatives in the universal choice set, different levels of sensitivity between pairs of alternatives, different numbers of variables used in the specification and their moment values, and the varying distributions of the response patterns to variables in the population.

The rest of the paper is organized as follows. Section 2 discusses the MNL and MMNL structures and the issues involved in sampling of alternatives. Section 3 describes the design of the numerical experiments. Section 4 presents the empirical results and discusses the important findings. The final section concludes the paper.

2. THE MODELS

2.1 The MNL Model (MNL)

The MNL model takes the following familiar form for the probability that individual q selects alternative i from the set of all available alternatives C.

$$P_{qi} = \frac{e^{\beta' X_{qi}}}{\sum_{j \in C} e^{\beta' X_{qj}}} \tag{1}$$

where X_{qi} is a vector of observed variables specific to individual q and alternative i, and β is a corresponding fixed parameter vector of coefficients.

Now, consider that the analyst decides to use only a subset of alternatives, D_q, for individual q. Let $\pi_q(D_q \mid i)$ be the probability under the researcher's selection mechanism of choosing subset D_q given that alternative i is chosen by individual q. For estimation purposes, D_q should include the chosen alternative, so that $\pi_q(D_q \mid i) = 0$ for any D_q that does not include i. The conditional probability of individual q choosing alternative i conditional on the researcher sampling the subset D_q for the individual may be derived in a straightforward manner using Bayes theorem as (*6*, p.68):

$$P_q(i \mid D_q) = \frac{P_{qi}\pi_q(D_q \mid i)}{\sum_{j \in C} P_{qj}\pi_q(D_q \mid j)} = \frac{P_{qi}\pi_q(D_q \mid i)}{\sum_{j \in D_q} P_{qj}\pi_q(D_q \mid j)} \tag{2}$$

The simplification in the denominator on the right side in the equation above is based on the fact that $\pi_q(D_q \mid j) = 0$ for j not in D_q. Next, for the MNL model, we can use Equation (1) in Equation (2) to write:

$$P_q(i \mid D_q) = \frac{e^{\beta' X_{qi}}\pi_q(D_q \mid i)}{\sum_{j \in D_q} e^{\beta' X_{qj}}\pi_q(D_q \mid j)} = \frac{e^{\beta' X_{qi} + \ln \pi_q(D_q \mid i)}}{\sum_{j \in D_q} e^{\beta' X_{qj} + \ln \pi_q(D_q \mid j)}} \tag{3}$$

The simplification in going from Equation (2) to Equation (3) is based on the cancellation of the denominators of P_{qi} in the MNL model (this cancellation is also fundamentally responsible for the IIA property). The analyst can use Equation (3) with any sampling mechanism s/he chooses, and only has to incorporate an additional variable $\ln \pi_q(D_q \mid i)$ in the utility of each alternative. The coefficient on this variable is restricted to 1 during estimation, which is based on maximizing the following conditional likelihood function:

$$C\langle(\beta) = \sum_q \sum_{i \in D_q} y_{qi} \ln P_q(i \mid \beta, D_q) \tag{4}$$

McFadden (*3*) proves that maximizing the above function provides consistent estimates of β. In the typical case when the analyst uses a random sampling approach, the following uniform conditioning property holds:

$$\pi_q(D_q \mid i) = \pi_q(D_q \mid j) \qquad \forall i, j \in D_q \tag{5}$$

Using this uniform conditioning property, Equation (3) collapses to a standard logit model with a choice set D_q (a subset of C) for individual q. Thus, a random sampling of alternatives allows consistent parameter estimation in the standard multinomial logit model.

2.2 The Mixed Multinomial Logit Model (MMNL)

The MMNL model is a generalization of the multinomial logit (MNL) model. Specifically, it involves the integration of the MNL formula over the distribution of random parameters. It takes the structure shown below:

$$P_{qi} = \int_{-\infty}^{+\infty} L_{qi}(\beta) f(\beta \mid \theta) d\beta \,,$$

where $L_{qi} = \dfrac{e^{\beta' X_{qi}}}{\sum_{j \in C} e^{\beta' X_{qj}}}$.

(6)

The use of the expression above in Equation (2) for the conditional probability of choosing alternative i given subset D_q immediately indicates that there is no simplification when sampling alternatives for the MMNL as for the MNL model in Equation (3). The reason is that, for the MNL case, a cancellation of the denominators in the probability expression takes place, putting the conditional probability back into the form of a tractable MNL expression. No such simplification occurs for the non-MNL models, because even under the assumptions of a uniform conditioning sampling approach, Equation (2) simplifies only to:

$$P_q(i \mid D_q) = \frac{P_{qi}(\theta)}{\sum_{j \in D_q} P_{qj}(\theta)} \tag{7}$$

The equation above requires the probability of each alternative to be computed with respect to all alternatives in the choice set. Thus, no sampling strategy will work in the case of the MMNL model (and more generally, in the case of other non-MNL models too such as the GEV class of models). But, an approximation in Equation (6) simplifies the expression in Equation (7). Specifically, one can approximate L_{qi} in Equation (6) as:

$$L_{qi} \approx \frac{e^{\beta' X_{qi}}}{\left(\frac{N}{S}\right) * \sum_{j \in D_q} e^{\beta' X_{qj}}}, \tag{8}$$

where S is the number of alternatives in D_q (i.e., the number of sampled alternatives) and N is the number of alternatives in C (i.e., the number of alternatives in the universal choice set). The term (N/S) is a factor that expands the sum of the denominator from the sampled alternatives to the full choice set. Then, one can write:

$$P_{qi}(\theta) \approx \frac{S}{N} * \int_{-\infty}^{\infty} \left(\frac{e^{\beta' X_{qi}}}{\sum_{j \in D_q} e^{\beta' X_{qi}}} \right) f(\beta \mid \theta) d\beta . \tag{9}$$

Equation (7) then collapses to:

$$\begin{aligned} P_q(i \mid D_q) &= \left(\int_{-\infty}^{\infty} \frac{e^{\beta' X_{qi}}}{\sum_{k \in D_q} e^{\beta' X_{qk}}} f(\beta \mid \theta) d\beta \right) \Bigg/ \left(\sum_j \int_{-\infty}^{\infty} \frac{e^{\beta' X_{qj}}}{\sum_{k \in D_q} e^{\beta' X_{qk}}} f(\beta \mid \theta) d\beta \right) \\ &= \int_{-\infty}^{\infty} \frac{e^{\beta' X_{qj}}}{\sum_{k \in D_q} e^{\beta' X_{qk}}} f(\beta \mid \theta) d\beta \end{aligned} \tag{10}$$

The simplification above occurs because the denominator in the first expression of Equation (10) is equal to 1. Thus, with the approximation in (9), the conditional probability is put back into a simple MMNL expression within the set of sampled alternatives. Of course, the approximation in (9) is the reason for the simplification. In general, the expression on the right side of Equation (9) is not a consistent estimator of $P_{qi}(\theta)$. Further theoretical exploration of this approximation is an important area for future research. In the current paper, we empirically test the ability to recover the underlying parameters and other relevant statistics using an MMNL model with a sample of alternatives and the expression in Equation (10).

3. EXPERIMENTAL DESIGN

In the numerical experiments of our study, we generate two datasets, one for the multinomial logit model and the other for the mixed multinomial logit model. Each dataset includes five independent variables for 200 alternatives for each of 750 observations. The values of the five independent variables for each of the 200 alternatives are drawn from a standard normal univariate distribution with the variables of the first 100 alternatives having a mean of 1 and the variables of the other 100 alternatives having a mean of 0.5.

For the multinomial logit dataset, the coefficients applied to each independent variable for each observation is taken as 1. The deterministic component of the utility is then calculated. The error term for each alternative and each observation is drawn independently from a type I extreme value distribution. This is achieved by obtaining draws from the uniform random distribution and applying the transformation -ln(-ln(u)) where u is a random number drawn from the uniform distribution between 0 and 1. The deterministic and the probabilistic components of the utilities for each alternative and each observation are added next to obtain the total utility for

each alternative. Finally, for each observation, the alternative with the highest utility is identified as the chosen alternative.

The steps involved in the generation of the dataset for the MMNL model are very similar to those used in generating the dataset for the MNL case. The only difference is that two of the five independent variables are assumed to have random coefficients. The random coefficients are assumed to be distributed univariate normal. As for the MNL data generation, the mean of the coefficients on all five independent variables is taken as 1. However, for two of these coefficients, we allow randomness across observations by drawing the coefficient from a univariate normal distribution with a mean value of 1 and a variance of 1 (this is, of course, achieved by drawing from a standard univariate normal distribution and adding 1). The error terms for the utilities are calculated in the same way as the MNL model, and the alternative with the highest utility is identified as the chosen alternative.

4. COMPUTATIONAL RESULTS

4.1 Estimation Issues

All the models were estimated using the GAUSS matrix programming language. The log-likelihood function and the gradient function for both the MNL and MMNL structures were coded. The Halton sequences required to simulate the probabilities in the mixed multinomial logit case were also generated using GAUSS.

In the first set of estimations involving the MNL model, the coefficients on the five independent variables in the simulated dataset were first estimated considering the full choice set of 200 alternatives. These results served as the benchmark to evaluate the performance of the random sampling of alternatives procedure. Next, we considered 6 different sample sizes for the number of alternatives in the random sampling: 5, 10, 25, 50, 100, and 150. For each size, the sampling was achieved through a GAUSS code that, for each observation, randomly selected (M-1) alternatives (without replacement) from the full choice set except the chosen alternative, and then added the chosen alternative to achieve the desired size M. Further, for each sample size, the sampling procedure just discussed was repeated 10 times using different random seeds to estimate the variance due to the sampling of alternatives.

In the second set of estimations involving the MMNL model, the same procedure as for the MNL was used in sampling alternatives. Unlike the MNL model, however, the maximum likelihood estimation of the MMNL model requires the evaluation of an analytically-intractable integral. The estimation is accomplished through a maximum simulated likelihood (MSL) approach using scrambled Halton draws with primes of 2 and 3 as the bases for the sequences (*7*). An important issue here is the number of Halton draws to use per observation. It is critical that the two-dimensional integral in the probability expressions of the MMNL model be evaluated accurately, so that the difference in model parameters between using a sample of alternatives and the full choice set can be attributed solely to the sampling of alternatives. In our MSL estimation of the MMNL model, we used 200 scrambled Halton draws based on extensive testing with different numbers of scrambled Halton draws. Specifically, we estimated an MMNL model using the MMNL dataset with 5 randomly sampled alternatives and the full choice set to represent the range of sample sizes of alternatives used in the experiments. For each of these two estimations, we estimated the model with different numbers of Halton draws, and found that the model parameters were basically indistinguishable beyond 200 Halton draws.

4.2 Evaluation Criteria

The focus of the evaluation effort is to assess the performance of the models estimated with a sample of alternatives relative to the model estimated with the full choice set. This evaluation was based on four criteria: (a) Ability to recover model parameters, (b) Ability to estimate the overall log-likelihood function accurately, (c) Ability to replicate the choice probability of the chosen alternative for each observation (i.e., ability to reproduce the individual likelihood function values), and (d) Ability to reproduce the aggregate shares of the alternatives. For the evaluation based on the latter three criteria, we applied the estimated parameter values from each estimation to the full choice set to compute the estimated choice probabilities for each of the 200 alternatives for each observation. The relevant values for the three criteria are then based on comparing the performance of each number of sampled alternatives on the full choice set with the true values computed from model estimation using the full choice set. This procedure brings the estimations with different sample sizes to a common platform and enables meaningful comparisons of model performance.

For each of the four criteria identified above, the evaluation of the proximity of the estimated and true values was based on two performance measures: (a) Root mean square error and (b) Mean absolute percentage error. Further, for each criterion-performance measure combination, we computed three properties: (a) Bias, or the difference between the mean of estimates for each sample size of alternatives across the 10 runs and the true values, (b) Simulation variance, or the variance in the relevant parameters across the 10 runs for each sample size of alternatives, and (c) Total error, or the difference between the estimated and the true values across all 10 runs for each sample size of alternatives. The performance statistics were compared across the different sample sizes to understand the effect of random sampling in each of the two model structures (MNL and MMNL), and across the two model structures to understand the differences of the effects of random sampling between them.

4.3 Performance Results

Tables 2 through 5 present the computational results. In each table, the error measures decrease in magnitude as the sample size increases, except for some minor aberrations in the bias measure for small sample sizes. Further, in each table, the error measures are larger for the MMNL model compared to the MNL model for each sample size of alternatives. This is to be expected, because of the theoretical result that random sampling is consistent in the MNL case while no such result holds for the MMNL case. We next discuss the important results from each table in turn.

Table 2 provides the measures of the ability to recover the model parameters. Several observations may be made from the table. First, for the MNL model, a doubling of the sample size of alternatives reduces the RMSE by about a fourth for sample sizes less than 50 (except for the decrease between 25 and 50 alternatives), and reduces the RMSE by about a half beyond sample sizes of 50. Similarly, for the MNL model, a doubling of the sample size of alternatives reduces the MAPE by about a third to a half as the sample size of alternatives is doubled, with the improvements in performance being steeper at higher sample sizes. These patterns are reflected in the bias, variance, and total error measures. Second, for the MMNL model, a doubling of the sample size reduces the bias and the total error of both the RMSE and MAPE by a fourth to a half up to a sample size of 50, but reduces these measures by half or more beyond a sample size of 50. The reduction in simulation variance in the MMNL model due to increasing sample size is more dramatic than the reduction in bias and total error measures and is rather consistent with a 50% reduction or more for a doubling in sample size (except for the results

corresponding to an increase from 50 to 100 alternatives). Third, the error measures for the MMNL model at a sample size of 5 are about 300-450% higher than for the MNL. However, the disparity between the MNL and the MMNL measures reduces to about 160-250% for sample sizes of 10, 25, and 50, and reduces further to about 40% beyond a sample size of 50. Further, the MAPE error is very high in the MMNL model for small sizes, and reduces quite substantially at higher sample sizes. Overall, the results suggest that it is very important to use high sample sizes for the MMNL model; at least a quarter of the full choice set and ideally a half of the full choice set or more.

Table 3 provides the results for the overall log-likelihood function value. The table shows that the MMNL RMSE error measures are large at small sample sizes in both absolute terms and relative to the MNL error measures. However, at sample sizes of 100 alternatives or more, the error measures become comparable between the MNL and MMNL models. The RMSE bias in this table is negative because the overall data fit from applying the parameters estimated from the sampling of alternatives procedure to the full set of alternatives can only be worse than the fit obtained by using the full choice set in estimation (which is the true convergent log-likelihood function value). Also, the MAPE bias and error measures in Table 2 are exactly the same because the log-likelihood is overestimated in magnitude (relative to the true value) by each of the 10 runs for each sample size.

Table 4 indicates that the MNL error measures are quite substantial at very small sample sizes for the individual log-likelihood function values (i.e. the probability of the chosen alternative). For example, the mean absolute percentage error is about 12% when a sample size of 5 is considered in the MNL. The percentage error is, as expected, much higher for the MMNL at small sample sizes. But, at larger sample sizes, it is remarkable that the MMNL error measures are very comparable to those from the MNL. Again, the results show that a sample size of 100 or more alternatives (or half the full choice set of alternatives or more) in the MMNL model provides good accuracy.

Table 5 mirrors the results from the earlier tables. The values in this table for the RMSE are smaller than for the other tables because the shares are computed at an aggregate level. The MAPE measure provides a better perspective here.

To summarize, three important observations may be drawn from Tables 2 through 5. First, random sampling of alternatives provides good estimates even for small sample sizes (i.e., small number of randomly sampled alternatives) in the MNL model. However, there is either a constant or increasing returns to scale in terms of accuracy and precision as the sample size is increased in the MNL model. Consequently, for the MNL model, the analyst would do well not to settle for very small sample sizes. Our results suggest a sample size of one-eighth of the full choice set as a minimum, and one-fourth of the full choice set as a good number of alternatives to target. Second, and as expected, the performance of the mixed multinomial logit model is very poor at small sample sizes. The good news for the analyst, however, is that the returns from increasing the sample size are much more dramatic in the MMNL model compared to the MNL model. In fact, at very high sample sizes, the accuracy of random sampling is comparable to the accuracy from the MNL model. As overall guidance, our recommendation based on the results would be that the analyst consider a sample size no less than a fourth of the full choice set and preferably half or more of the full choice set. The reader will note that even using half of the full choice set, though computationally expensive, can still lead to quite considerable savings in computational time compared to using the full choice set in the MMNL model. On the other hand, using a very small sample size may be good for computational time, but is literally

"garbage" from an accuracy perspective. Third, a comparison of the MAPE values from Tables 2 and 3 show that the overall log-likelihood value is more accurately estimated than the parameter values, especially for the MMNL model. This suggests a rather flat log-likelihood function near the optimum; that is, closeness to the log-likelihood function does not necessarily imply closeness in model parameters too.

In order to understand the effect of random sampling for unrestricted choice sets of smaller sizes than 200, we also generated datasets with 100 alternatives and 50 alternatives and undertook the same kind of analysis as just discussed above for the case with 200 alternatives. Tables 6 through 9 present the simulation results for datasets with 100 alternatives (we are not presenting the tables for the simulation results with 50 alternatives due to space constraints). The patterns of the results are similar to the case with 200 alternatives, and the same overall conclusions may be drawn.

5. SUMMARY AND CONCLUSIONS

Many activity and travel choice decisions made by individuals involve a large number of choice alternatives. Examples include activity participation location choice, residential location choice, and route choice. McFadden (*3*) proved that, if the analyst is willing to assume a simple multinomial logit (MNL) formulation for the behavior underlying the choice process, a sampling of alternatives scheme will provide consistent model parameters. Several researchers have exploited this result for estimations in different empirical contexts (*8-12*). However, there has been no systematic study (until this paper) of the effect of sample size on the empirical accuracy and efficiency of the estimated parameters. Further, with recent advances in the field, researchers are increasingly turning to more behaviorally realistic discrete choice models in analysis for which McFadden's (*3*) result does not hold. At the same time, sampling of alternatives can reduce computational time quite substantially compared to using the full choice set in these advanced models. Thus, it is of value to study the effect of sample size on model performance in these non-MNL models. This paper develops an evaluation framework for examining the effect of the sample size of alternatives on model performance both in an MNL context and a mixed multinomial logit (MMNL) context.

The results from this paper show the good numerical performance of the MNL model even with very small sample sizes. However, since the "bang for the buck" is high as one proceeds to larger sample sizes of alternatives, it is advisable to consider sample sizes that are not too small. Based on our results, we recommend the use of an eighth of the size of the full choice set as a minimum, and suggest a fourth of the full choice set as a desirable target. The performance of the MMNL model, on the other hand, is very poor at small sample sizes. However, the "bang for buck" is even better for the MMNL model with increasing sample sizes compared to the MNL case. At a minimum, we suggest using a fourth of the full choice set. However, we strongly suggest using one-half of the full choice set or more based on the numerical exercises in this paper.

As with any numerical exercise, the usual cautions for generalizing the results apply to this paper too. There is certainly a need for more computational and empirical research on the topic of sampling of alternatives in different settings (such as different patterns of correlation among exogenous variables, different levels of sensitivity and heterogeneity in the sensitivity to variables, different numbers of variables with random coefficients, and different numbers of decision makers in the sample) to draw more definitive conclusions. In the meantime, the results

of this paper should serve as a good guide to the analyst faced with modeling choice situations with large choice sets.

A final note before closing. This paper should not be viewed as encouraging sampling of alternatives in a non-MNL setting. In such settings, it is always most ideal to consider the full choice set. But, if considering the full choice set is difficult for the MNL model, it is substantially more difficult for the non-MNL models. The purpose of this numerical analysis is to provide some guidance to analysts wanting to use non-MNL models, but are simply unable to consider the full choice set. The results here should be viewed as an effort to set minimum sample size guidelines for the MMNL model when the full choice set cannot be considered.

ACKNOWLEDGEMENT

The comments from three anonymous referees are greatly appreciated.

REFERENCES

1. Swait, J. Choice Set Generation Within the Generalized Extreme Value Family of Discrete Choice Models, *Transportation Research Part B,* Vol. 35, No. 7, 2001, pp. 643-666.

2. Basar, G., and C.R. Bhat. A Parameterized Consideration Set Model for Airport Choice: An Application to the San Francisco bay Area. *Transportation Research Part B*, forthcoming, 2004.

3. McFadden, D. Modeling the Choice of Residential Location. In *Transportation Research Record 673*, TRB, National Research Council, Washington D.C., 1978, pp. 72-77.

4. Bhat, C.R., and J.Y. Guo. A Mixed Spatially Correlated Model: Formulation and Application to Residential Choice Modeling. *Transportation Research Part B*, Vol. 38, No. 2, 2004, pp. 147-168.

5. Bekhor, S., Ben-Akiva, M., and M.S. Ramming. Adaptation of Logit Kernel to Route Choice Situation. In *Transportation Research Record 1805*, TRB, National Research Council, Washington D.C., 2002, pp. 78-85.

6. Train, K. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, UK, 2003.

7. Bhat, C.R. Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences. *Transportation Research Part B,* Vol. 37, No. 9, 2003, pp. 837-855.

8. Sermons, M.W., and F.S. Koppelman. Representing Differences Between Female and Male Commute Behavior in Residential Location Choice Models. *Journal of Transport Geography*, Vol. 9, No. 2, 2001, pp. 101-110.

9. Waddell, P. Accessibility and Residential Location: The Interaction of Workplace, Residential Mobility, Tenure, and Location Choices. Presented at the *Lincoln Land Institute TRED Conference*, 1996 (http://www.odot.state.or.us/tddtpau/modeling.html).

10. Bhat, C.R., Govindarajan, A., and V. Pulugurtha. Disaggregate Attraction-End Choice modeling: Formulation and Empirical Analysis. In *Transportation Research Record 1645*, TRB, National Research Council, Washington D.C., 1998, pp. 60-68.

11. Guo, J.Y., and C.R. Bhat. Residential Location Choice Modeling: A Multinomial Logit Approach. Technical Paper, Department of Civil Engineering, The University of Texas at Austin, 2001.

12. Ben-Akiva, M.E., and J.L. Bowman. Integration of an Activity-Based Model System and a Residential Location Model. *Urban Studies*, Vol. 35, No. 7, 1998, pp. 1131-1153.

13. Srour, I.M., Kockelman, K.M., and T.P. Dunn. Accessibility Indices: A Connection to Residential Land Prices and Location Choices. In *Transportation Research Record 1805,* TRB, National Research Council, Washington D.C., 2002, pp. 25-34.

14. Schlich, R., Simma, A., and K.W. Axhausen. Destination Choice Modeling for Different Leisure Activities. *2nd Swiss Transport Research Conference 2002,* Ascona, March 2002.

15. Pozsgay, M.A., and C.R. Bhat. Destination Choice Modeling for Home-Based Recreational Trips: Analysis and Implications for Land-Use, Transportation, and Air Quality Planning. In *Transportation Research Record 1777,* TRB, National Research Council, Washington D.C., 2002, pp. 47-54.

LIST OF TABLES

TABLE 1 Earlier Studies Using a Subset of Choice Alternatives

Study	Choice Modeled	Alternatives Considered (Total Number)	Sampling Mechanism	Model Structure
Guo and Bhat *(11)*	Residential Location	TAP Zones (900)	Simple Random Sampling	MNL
Sermons and Koppelman *(8)*	Residential Location	Census Tracts (1099)	Simple Random Sampling	MNL
Ben-Akiva and Bowman *(12)*	Residential Location	TAZ (787)	Stratified Importance Sampling	Nested Logit
Bekhor, S. *et al. (5)*	Route Choice	Routes (50)	Simple Random Sampling	Kernel Logit
Waddell *(9)*	Residential Location	TAZ (761)	Simple Random Sampling	Nested Logit
Schlich *et al. (14)*	Destination Choice	Municipal Level (555)	Simple Random Sampling	MNL
Pozsgay and Bhat *(15)*	Destination Choice	TSZ (919)	Simple Random Sampling	MNL
Bhat *et al. (10)*	Attraction-end Choice	TAZ (858)	Importance Sampling	MNL

TABLE 2 Evaluation of Ability to Recover Model Parameters (with a choice set of 200 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample					
			5	10	25	50	100	150
MNL	RMSE	Bias	0.0323	0.0226	0.0249	0.0194	0.0086	0.0050
		Simulation Variance	0.0658	0.0488	0.0250	0.0233	0.0152	0.0075
		Total Error	0.0733	0.0538	0.0353	0.0303	0.0175	0.0090
	MAPE	Bias	3.1002	2.1288	2.3898	1.6856	0.8086	0.3838
		Simulation Variance	5.4227	3.7539	2.0599	1.8808	1.3079	0.6065
		Total Error	6.0238	4.2848	2.9142	2.4236	1.5140	0.7713
MMNL	RMSE	Bias	0.3480	0.1587	0.1017	0.0768	0.0184	0.0080
		Simulation Variance	0.2059	0.0893	0.0464	0.0233	0.0185	0.0099
		Total Error	0.4043	0.1821	0.1118	0.0802	0.0261	0.0127
	MAPE	Bias	23.8286	14.2235	9.3792	6.8456	1.6836	0.7006
		Simulation Variance	31.4289	7.3227	3.4461	1.7203	1.4214	0.7662
		Total Error	24.6240	14.8463	10.0229	7.3630	2.1272	1.0454

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error

TABLE 3 Evaluation of Ability to Estimate Overall Log-Likelihood Function Value (with a choice set of 200 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample					
			5	10	25	50	100	150
MNL	RMSE[1]	Bias	-8.2262	-4.2624	-1.8495	-1.2370	-0.4206	-0.1177
		Simulation Variance	3.7895	3.4157	0.9942	0.8063	0.3075	0.0570
		Total Error	9.0571	5.4622	2.0998	1.4766	0.5210	0.1308
	MAPE[2]	Bias	0.3591	0.1861	0.0807	0.0540	0.0184	0.0051
		Simulation Variance	0.1417	0.1204	0.0369	0.0302	0.0102	0.0015
		Total Error	0.3591	0.1861	0.0807	0.0540	0.0184	0.0051
MMNL	RMSE[1]	Bias	-67.7402	-41.4890	-15.7482	-7.3515	-0.6999	-0.1324
		Simulation Variance	16.2127	13.1415	4.6363	0.7306	0.4039	0.0450
		Total Error	69.6533	43.5206	16.4165	7.3877	0.8081	0.1398
	MAPE[2]	Bias	2.7927	1.7104	0.6492	0.3031	0.0289	0.0055
		Simulation Variance	0.4989	0.4361	0.1562	0.0248	0.0153	0.0015
		Total Error	2.7927	1.7104	0.6492	0.3031	0.0289	0.0055

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error

TABLE 4 Evaluation of Ability to Estimate Individual Choice Probabilities (with a choice set of 200 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample					
			5	10	25	50	100	150
MNL	RMSE[1]	Bias	0.0100	0.0065	0.0069	0.0056	0.0030	0.0017
		Simulation Variance	0.0199	0.0146	0.0074	0.0073	0.0049	0.0023
		Total Error	0.0223	0.0160	0.0101	0.0092	0.0058	0.0028
	MAPE[2]	Bias	5.4048	3.8059	3.9846	2.8590	1.3295	0.7826
		Simulation Variance	10.1145	7.0496	3.8917	3.4659	2.2394	1.0874
		Total Error	11.5424	7.9121	5.4241	4.3606	2.5686	1.3812
MMNL	RMSE[1]	Bias	0.0450	0.0410	0.0278	0.0201	0.0048	0.0015
		Simulation Variance	0.0174	0.0157	0.0081	0.0037	0.0032	0.0015
		Total Error	0.0483	0.0439	0.0289	0.0204	0.0058	0.0021
	MAPE[2]	Bias	28.0816	22.4440	14.9737	10.8548	2.6249	0.9960
		Simulation Variance	11.4763	10.6652	5.1676	2.4932	2.1856	1.0523
		Total Error	29.9596	24.2285	15.6463	11.0900	3.2574	1.4390

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error

TABLE 5 Evaluation of Ability to Estimate Aggregate Shares of Alternatives (with a choice set of 200 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample					
			5	10	25	50	100	150
MNL	RMSE[1]	Bias	5.4040E-05	3.5401E-05	4.1712E-05	3.5245E-05	1.6883E-05	9.1212E-06
		Simulation Variance	1.0761E-04	8.1337E-05	4.1007E-05	4.1110E-05	2.6953E-05	1.3054E-05
		Total Error	1.2042E-04	8.8707E-05	5.8493E-05	5.4151E-05	3.1804E-05	1.5925E-05
	MAPE[2]	Bias	1.3659	0.7076	1.1177	1.2655	0.7648	0.3656
		Simulation Variance	2.9415	2.2135	0.9772	1.2514	0.8804	0.3892
		Total Error	3.3484	2.3569	1.4558	1.7745	1.0421	0.5217
MMNL	RMSE[1]	Bias	3.4953E-04	3.1941E-04	2.2537E-04	1.6793E-04	4.0930E-05	1.3208E-05
		Simulation Variance	1.2847E-04	9.7856E-05	5.5666E-05	2.4174E-05	2.3705E-05	1.1346E-05
		Total Error	3.7239E-04	3.3407E-04	2.3214E-04	1.6967E-04	4.7299E-05	1.7412E-05
	MAPE[2]	Bias	11.6474	12.2540	9.1110	6.9274	1.7033	0.5252
		Simulation Variance	3.5642	2.7321	1.5928	0.6102	0.7976	0.3325
		Total Error	11.8229	12.3598	9.1469	6.9371	1.7284	0.5850

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error

TABLE 6 Evaluation of Ability to Recover Model Parameters (with a choice set of 100 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample				
			5	10	25	50	75
MNL	RMSE[1]	Bias	0.0558	0.0510	0.0438	0.0100	0.0052
		Simulation Variance	0.0667	0.0371	0.0262	0.0182	0.0083
		Total Error	0.0860	0.0630	0.0495	0.0221	0.0101
	MAPE[2]	Bias	5.0302	4.5650	3.9828	0.8414	0.3825
		Simulation Variance	4.9779	2.9056	1.7038	1.6036	0.7149
		Total Error	6.9460	5.2267	4.2026	1.7572	0.8121
MMNL	RMSE[1]	Bias	0.2370	0.2144	0.1624	0.0437	0.0064
		Simulation Variance	0.1996	0.1562	0.0863	0.0628	0.0085
		Total Error	0.3098	0.2583	0.1028	0.0619	0.0153
	MAPE[2]	Bias	19.3345	16.7802	7.2842	1.1252	0.3997
		Simulation Variance	14.4498	9.6755	3.4824	1.9284	0.7263
		Total Error	20.9486	17.8954	8.2741	2.1927	0.8462

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error

TABLE 7 Evaluation of Ability to Estimate Overall Log-Likelihood Function Value (with a choice set of 100 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample				
			5	10	25	50	75
MNL	RMSE[1]	Bias	-8.1384	-4.4595	-2.4939	-0.6010	-0.13653
		Simulation Variance	4.6587	1.9117	0.9511	0.3073	0.0844
		Total Error	9.3774	4.8520	2.6691	0.6750	0.1605
	MAPE[2]	Bias	0.4295	0.2354	0.1316	0.0317	0.0072
		Simulation Variance	0.2038	0.0799	0.0408	0.0154	0.0036
		Total Error	0.4295	0.2354	0.1316	0.0317	0.0072
MMNL	RMSE[1]	Bias	-55.9366	-28.5588	-10.3647	-0.8245	-0.1673
		Simulation Variance	15.2852	8.1165	1.3525	0.4824	0.0926
		Total Error	57.9874	29.2065	13.2476	1.2460	0.1836
	MAPE[2]	Bias	2.7778	1.4182	0.5353	0.0436	0.0012
		Simulation Variance	0.6884	0.2871	0.1241	0.0356	0.0036
		Total Error	2.7778	1.4182	0.5353	0.0436	0.0012

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error

TABLE 8 Evaluation of Ability to Estimate Individual Choice Probabilities (with a choice set of 100 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample				
			5	10	25	50	75
MNL	RMSE[1]	Bias	0.0172	0.0160	0.0139	0.0032	0.0017
		Simulation Variance	0.0218	0.0122	0.0076	0.0065	0.0028
		Total Error	0.0277	0.0201	0.0158	0.0072	0.0033
	MAPE[2]	Bias	6.9878	6.4866	5.4740	1.4638	0.8224
		Simulation Variance	8.9242	5.2463	3.2561	2.5962	1.1896
		Total Error	11.0300	8.3583	6.3631	3.0116	1.4254
MMNL	RMSE[1]	Bias	0.0587	0.0417	0.0267	0.0083	0.0019
		Simulation Variance	0.0209	0.0149	0.0892	0.0099	0.0031
		Total Error	0.0623	0.0443	0.0272	0.0098	0.0037
	MAPE[2]	Bias	25.3416	18.1166	8.2812	2.0383	0.8562
		Simulation Variance	11.7932	8.7375	4.7421	3.1923	1.1899
		Total Error	27.3515	19.6160	11.7294	4.2853	1.6723

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error

TABLE 9 Evaluation of Ability to Estimate Aggregate Shares of Alternatives (with a choice set of 100 alternatives)

Model	Performance Measure	Estimator Property	Number of alternatives considered in the random sample				
			5	10	25	50	75
MNL	RMSE[1]	Bias	1.4643E-04	1.3196E-04	1.2070E-04	2.3040E-05	1.1947E-05
		Simulation Variance	1.7681E-04	8.7414E-05	5.6720E-05	5.3854E-05	2.1934E-05
		Total Error	2.2957E-04	1.5829E-04	1.3336E-04	5.8575E-05	2.4976E-05
	MAPE[2]	Bias	3.5782	3.2278	2.9787	0.2510	0.1355
		Simulation Variance	3.1723	1.1710	0.7622	0.9567	0.3570
		Total Error	4.1769	3.3214	3.0072	1.0061	0.3792
MMNL	RMSE[1]	Bias	6.5060E-04	4.8618E-04	2.8618E-04	4.8855E-05	2.1618E-05
		Simulation Variance	1.7714E-04	1.2274E-04	8.2274E-05	8.2810E-05	2.3810E-05
		Total Error	6.7429E-04	5.0143E-04	3.1432E-04	6.2173E-05	2.4322E-05
	MAPE[2]	Bias	12.5711	10.0086	6.0086	1.5332	0.1432
		Simulation Variance	3.0459	1.8796	1.0880	0.9038	0.3779
		Total Error	16.3075	11.0118	6.0118	2.1797	0.4296

[1]RMSE: Root Mean Square Error

[2]MAPE: Mean Absolute Percentage Error