MIDDLE STATES GEOGRAPHER - VOL. 28, 1995

# CHOICE SET DEFINITION IN SHOPPING DESTINATION CHOICE MODELS: SOME SENSITIVITY ANALYSIS RESULTS

Pasquale A. Pellegrini
Department of Geography and
National Center for Geographic Information and Analysis,
Wilkeson Quad, State University of New York at Buffalo
Buffalo, NY 14261

**ABSTRACT** It is widely recognized that defining spatial choice sets is difficult primarily because of the large number of alternatives usually associated with spatial decision making. In this paper, the Multinomial Logit discrete choice model is calibrated with random sub-sets of shopping destinations in order to explore parameter sensitivity to choice set mis-specification. The results presented here are derived from a sensitivity analysis based on a combinatoric choice sub-set generation procedure that systematically reduces the original choice set and re-estimates the model parameters. The results indicate that differences in distance deterrence across income groups holds true in reduced choice set situations, indicating greater distance deterrence amongst high income individuals over low income consumers. In addition, the differential importance of chain image across racial groups displays consistency throughout the sensitivity analysis providing evidence that this variable is just as important as standard variables like store size and competition in store choice models.

## INTRODUCTION

The difficulty involved in properly defining the choice set faced by decison makers in spatial choice contexts increases the likelihood for model mis-specification and erroneous parameter estimates (Thill, 1992; Pellegrini and Fotheringham, 1994). The choice set refers to the group of discrete alternatives faced by an individual in the decision making process. The specification of each individual's choice set is complicated by the sheer number of choice alternatives available in spatial decision situations like the choice of grocery store or residential location. In contrast, aspatial choice situations, such as travel mode choice, typically contain only a handful of alternatives (Fotheringham and O'Kelly, 1989). Very often, however, the choice set definition problem is avoided by analysts by making the bold assumption that they are capable of defining one choice set from which each individual evaluates and forms a destination choice. Of course, it is far more reasonable to make the assumption that individuals differ in their perceived choice sets and their knowledge of available alternatives, but much more difficult to model.

This paper presents a discussion of the results from an empirical analysis designed to examine parameter sensitivity to spatial choice set mis-specification in a disaggregate shopping destination choice analysis. A full discussion of the relevant literature, data and methodology, and results from this sensitivity analysis is available in Pellegrini, Fotheringham and Lin (1994), but this paper provides a summary of the major research findings. In essence, spatial choice set mis-specification may lead to incorrect parameter estimation of the utility function for the decision makers, and subsequently, erroneous interpretations of individual behavior.

## PREVIOUS CHOICE SET SPECIFICATION RESEARCH

Two mis-specification scenarios are likely to arise in regards to the analyst assigning an arbitrary choice set in a spatial choice modeling application. First, the choice set assigned by the analyst can be in the individual's *true* choice set. In this case, the model parameters can still be estimated consistently and choice probabilities correctly predicted by a random utility based spatial choice model (McFadden, 1978). In contrast, erroneous model parameter estimates are expected from the case where the choice set defined by the analyst includes alternatives actually never evaluated by the decision maker (Williams and Ortuzar, 1982). Here, the choice model assigns non-negative probabilities to all alternatives in the choice set, including those that are not in the true choice set. This results in inconsistent estimates of the choice function, and faulty interpretations of individual behavior - the most serious consequence of choice set mis-specification (Thill, 1992). The main concern of the research reported here is this latter scenario, and the stability of parameter estimates to choice set mis-specification is studied using a random combinatoric choice sub-set generation procedure adapted from Manski (1977).

A fair amount of research directed at better defining choice sets exists, especially in regards to the behavioral foundations of *choice set generation*, although most are in aspatial rather than spatial choice contexts, and are reviewed in Thill (1992) and in the context of retailing by Pellegrini and Fotheringham (1994). Briefly, this work is based on the notion of what Manski (1977) introduced as the two-stage paradigm of discrete choice analysis, where the first stage is choice set generation and the second is the actual choice generation. In the choice set generation stage, constraints limiting the set of available alternatives to the decision-maker are identified and used to help establish the feasible sub-set of alternatives that the decision-maker actually considers for choice. However, the two-stage choice modelling framework is of little use for practitioners interested in spatial applications where choice sub-sets are spatially dependent, the alternatives and decision-makers are distributed and interrelated over space, and the number of feasible alternatives is frequently large (Fotheringham, 1988; Thill, 1992). Specifically, these formulations are limited when the individual's universal choice set can not be decomposed into sub-sets given the spatial distribution of destinations as in a store choice situation.

In view of the difficulties involved in choice set definition, in terms of choice set generation models, the effects of constraints on choice, and the potentially detrimental consequences of choice set mis-specification, this paper attempts to empirically evaluate parameter sensitivity to choice set definition by estimating a series of destination choice models on random sub-sets of the complete choice set used for an empirical store choice study. In this way, we avoid the difficulties of defining choice sets amongst individuals with constraints, whilst having greater confidence in variable parameters that are consistent across choice sets, and viceversa with unstable parameter estimates. The *competing destinations* spatial choice model, a variation of the multinomial logit (MNL) model, is used in this analysis (Fotheringham, 1986; 1988).

## DATA AND METHODOLOGY

The empirical focus of this sensitivity analysis is taken from Fotheringham and Trew's (1993) disaggregate analysis of chain image, and the effects of race and income, in Gainesville, Florida. In this paper, we consider the sensitivity of the parameter estimates derived in the Fotheringham and Trew study to variations in choice set definition. The original analysis is repeated by using the same fourteen store choice set, and the estimated parameters are used as the baseline figures to compare with parameter

estimates obtained from model calibrations with various sub-sets of the full choice set. Six models are calibrated, one using all the consumers and the other five only for consumers segmented by income and by race, although the models contain the same "core" variables described briefly below.

The store choice data used in Fotheringham and Trew (1993) includes the location, size, and chain membership of fourteen major supermarkets in Gainesville. This choice set was determined both by the actual store choices of the 432 consumers in the sample, and the minimum store size of 19,000 square feet. As with many other store choice studies, this choice set definition is a compromise between the extreme choice set possibilites - with one consisting of all the grocery stores in Gainesville, and at the other extreme, just those stores most frequently visited by the sample consumers. It is more likely, however, that the consumers considered some sub-set of the full fourteen store choice set, and store choice models calibrated using random samples of these sub-sets is the focus of this investigation.

Six MNL models specified in the competing destinations format are calibrated for all the consumers, and consumers segmented by income and race, in order to investigate the variation in store choice behavior across market segments and, subsequently, the stability of these parameter estimates to choice set mis-specification. The 432 sample consumers break down as follows: 97 high income (>$35,000), 165 medium income ($15,000-$35,000), 170 low income (<$15,000), 384 whites and 48 blacks. The first core variable is the spatial separation, or distance $d_{ij}$, of the supermarket $j$ from the residence of the consumer $i$. The distance parameter is expected to yield negative estimates given both the psychological and economic views of distance as a surrogate for information about a store and the increased real price of goods through additional transportation costs, respectively. Another variable common to all models is the size of a store, $S_j$, measured in square footage. This variable serves as a surrogate for the variety of products offered, and is expected to yield a positive parameter when the model is calibrated.

Store competition, $C_j$, is the third core variable in the model specifications. This variable defines the spatial competition at store $j$ by a simple potential measure:

$$C_j = \sum_{k \neq j} \frac{S_k}{\exp(\beta d_{jk})} , \qquad (3)$$

where $C_j$ is the spatial competition at $j$, $S_k$ is the size of store $k$, $d_{jk}$ is the distance from $j$ to $k$, and the parameter $\beta$ is set to 1.0 (Fotheringham and Trew, 1993). The rationale for adding this variable to the store choice MNL model, which forms the competing destinations model, relates to the removal of regularity and the independence from irrelavant alternatives properties from the MNL, and is discussed in a series of papers by Fotheringham (1983; 1986; 1988). The competition variable is also justified from both an economic and psychological point of view. From an economic standpoint, consumers may seek to minimize the distance to alternative stores to facilitate comparison shopping or have an alternative should the first store not have a certain product, resulting in a positive parameter estimate for store competiton. From a psychological perspective, some hierarchical information processing is assumed in spatial choice situations where individuals view alternatives in spatial clusters, and make choices between clusters, underestimating the number of alternatives in large clusters according to the well known psychophysical law (Stevens, 1957). This argument would produce a negative parameter estimate because consumers would be less likely to select an alternative in close proximity to other stores. Given this situation where both positive and negative parameters are feasible, it is possible that the two explanations can counteract each other's influence in the model.

The fourth variable is the neighborhood characteristic, $N_j$, defined as the proportion of individuals living in the neighborhood of store $j$ that are maximally dissimilar to the market segment of which individual $i$ is a member (Fotheringham, 1988). Consequently, for high income consumers, the store choice model specified with this variable defined as the proportion of low income consumers in the neighborhood of store $j$, and it is similarly defined for each income and race market segment.

The final variable in the store choice model is a chain image dummy variable designed to examine the influence of chain image on patronage behavior across market segments. The chain image dummy is set to 1 for 'upmarket' stores (Publix, Albertson's and Kash n Karry) and 0 for the 'downmarket' stores (Winn dixie and Food 4 Less). This binary definition for the chain dummy was used by Fotheringham and Trew (1993) to offset the colinearity problem between store size and chain image when the chain image dummy was defined to identify each chain. By defining this variable in the exponential form, one can assess how many times an individual is more or less likely to patronize a store belonging to an upmarket chain, *ceteris paribus*, by taking the exponential of the estimate.

**Model Specification and the Generation of Choice Sets**

The general store choice model calibrated for the three income groups, two race groups, and all of the respondents is specified as:

$$P_{ij} = \frac{d_{ij}^{\alpha_1} \cdot S_j^{\alpha_2} \cdot C_j^{\alpha_3} \cdot N_j^{\alpha_4} \cdot \exp(\alpha_5 I_j)}{\sum_k d_{ik}^{\alpha_1} \cdot S_k^{\alpha_2} \cdot C_k^{\alpha_3} \cdot N_k^{\alpha_4} \cdot \exp(\alpha_5 I_k)}, \quad (2)$$

where

$P_{ij}$ is the probability that consumer $i$ selects store $j$,

$d_{ij}$ is the distance from the consumer's residence to store $j$,

$S_j$ is the size of store $j$,

$C_j$ is the competition at store $j$,

$N_j$ is the neighborhood characteristic,

$I_j$ is the chain image dummy,

and $k$ is the size of the choice set. The $\alpha$s are parameters representing the relationship between $P_{ij}$ and the independent variables.

In terms of model diagnostics, the $\rho^2$ statistic is used to assess model fit to the data. It is calculated as follows:

$$\rho^2 = 1 - \frac{L(1)}{L(0)} \quad , \tag{5}$$

where $L(1)$ and $L(0)$ are the values of the log-likelihood function of the model containing the estimated parameters and the log-likelihood of the null model, respectively. This measure is also known as the 'equal shares' statistic, and values above 0.2 are indicative of a good fit.

The analysis involves the generation of a complete combinatoric inventory of possible choice sets ranging in size from thirteen to three stores using a computer algorithm that operates similar to the choice set generation procedures discussed by Manski (1977) and Williams and Ortuzar (1982). From this complete combinatoric inventory, the model in equation (2) is calibrated and the diagnostic test in equation (3) is calculated for the various choice sets. For the very large sets of combinations, a 10% random sample of the total combinations for each choice set size is selected for model calibration, since the computational burden grows geometrically with choice set size. The econometric software package LIMDEP (Greene, 1991) is used in all the logit model calibrations.

## SENSITIVITY ANALYSIS RESULTS

The estimated parameters of equation (2) for the original fourteen store choice set across all market segments are presented in table 1. These parameter estimates are used as the baseline figures for comparison with the results obtained from the random combinatoric sub-sets of alternatives. Distance from a consumer's residence is clearly the prime variable in explaining store choice. An interesting trend in the distance parameter estimates across the three income groups indicates that low-income consumers ($\alpha^1$=-1.65) are willing to travel longer distances than high-income consumers ($\alpha^1$=-2.17). Presumably, this behavior is related to a search for lower prices producing a flatter distance-decay function for grocery shopping in Gainsville.

The store size variable is positive and significant for low income consumers which is not surprising given the usual association between increased store size and lower prices. The competition variable is not significant for any of the income groups, indicating that perhaps the opposite influences of hierarchical choice and comparison shopping are cancelling each other out. The neighborhood characteristic variable is negative, as expected, but only significant for high income consumers (at 90% confidence level). This suggests that high income consumers are deterred from shopping in low income neighborhoods, but the medium and low income consumers are not similarly deterred from shopping in high income neighborhoods. In terms of chain image, a significant preference for the upmarket chains exists for all three income groups, and this preference ranges from the low income consumers being almost twice as likely to patronize a store belonging to an upmarket chain, whilst the high income market segment is five times as likely to patronize an upmarket chain, *ceteris paribus* (by taking the exponential of the estimates).

Table 1: Model parameter (or baseline) estimates for the fourteen store choice set.

| *Variable* | *Models* | *Income* | | | *Race* | |
|---|---|---|---|---|---|---|
| | All | Low | Medium | High | White | Black |
| Distance | -1.80 | -1.63 | -1.92 | -2.12 | -1.85 | -1.68 |
| | (20.63) | (12.64) | (12.40) | (9.28) | (18.91) | (6.48) |
| Store Size | -0.09 | 1.13 | 0.22 | -1.03 | 0.44 | 0.88 |
| | (0.32) | (2.35) | (0.45) | (1.33) | (1.28) | (0.83) |
| Competition | 0.05 | 0.07 | 0.03 | -0.06 | 0.05 | -0.02 |
| | (1.76) | (1.49) | (0.55) | (0.86) | (1.56) | (0.22) |
| Neighborhood | | -0.03 | -0.03 | -0.44 | -0.05 | 0.15 |
| | | (0.20) | (0.16) | (1.87) | (1.25) | (0.41) |
| Upmarket Dummy | | 0.62 | 0.77 | 1.63 | 1.19 | -0.80 |
| | | (3.09) | (3.45) | (3.80) | (7.33) | (2.21) |
| Log-likelihood | -843.06 | -328.67 | -313.91 | -166.65 | -712.73 | -91.60 |
| $\rho^2$ | 0.26 | 0.27 | 0.28 | 0.35 | 0.30 | 0.28 |

Note: Figures in parentheses are the *t*-statistics.

With regards to store choice and race, table 1 indicates that distance deterrence in grocery shopping is equally strong for white and black consumers. On the other hand, store size is a positive influence for both racial groups, although not to a significant level. The proximity of competitors appears to attract white consumers, but does not influence black consumers, although neither parameter is significant at the 90% confidence level. Similarly, the neighborhood characteristic of the selected store does not appear to deter either black or white consumers. The most interesting parameter associated with the racially defined market segments is the chain image dummy. Here, the results suggest that white consumers are almost three times as likely to choose an upmarket store, whilst black consumers are actually slightly deterred from choosing an upmarket store. In all the model calibrations, the $\rho^2$ values range from 0.26 to 0.35, indicating a good model fit to the data.

**Results with Reduced Choice Set Size**

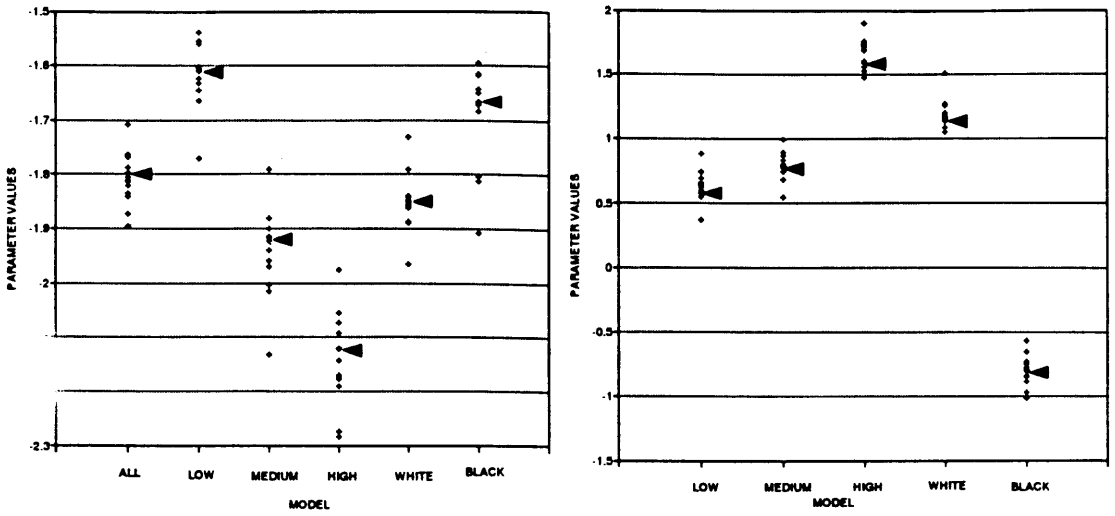
Given the above results, and those in Fotheringham and Trew (1993), indicating that shoppers do not necessarily shop at the closest store, and more generally, that important differences exist in store choice across different consumer groups, the sensitivity of the various parameter estimates to choice set mis-specification is tested. The baseline results with the fourteen store choice set indicates a trend in the distance deterrence amongst income groups, a relatively insignificant role of store size, competition, and neighborhood characteristic in store choice, but a strong influence of chain image in store choice for the racial groups.

The sensitivity analysis indicates parameter instability for store size, competition, and the neighborhood characteristic variables. A high level of dispersion around the baseline estimates, and parameter instability in terms of sign and significance for these three variables indicates their sensitivity to choice set specification. In addition, the variable parameter estimates become increasingly unstable with

decreasing choice set size. However, the all consumer model shows greater parameter stability, indicating the potential for biased parameter estimates by not disaggregating by market segment. Only the neighborhood characteristic parameter estimate for high income consumers remains significant and stable amongst choice set specifications suggesting economic prejudice in store choice for that market segment.

In contrast, the distance and chain image variables remained relatively stable. In figure 1a, the distance parameter for each thirteen store choice set is shown (indicated by addition symbols) along with the baseline parameter (indicated by arrowheads). The distance parameter shows little variation around the baseline parameter estimate, but the same general trend of increasing deterrence from low to high income groups remains. The robustness of the distance parameters continues to be very encouraging when the random samples of smaller choice sets are calibrated. In figures 2a and 2b, the distance parameters for the ten, seven and five alternative choice set calibrations are plotted as frequency distributions. The distributions of parameter values peak at the baseline results for the low and high income market segments, are symmetrical, have little deviation from the baseline parameter, and exhibit only sporadic occurrences of extreme values. Most important, however, is the consistency of the trend indicating greater distance deterrence amongst higher income individuals than their lower income counterparts. The robustness of the variation in the frictional effect of distance on spatial choice amongst income groups provides further support for the notion that low income consumers may be more willing to travel longer distances than high income consumers for grocery shopping, perhaps in search of lower prices.

(a) (b)



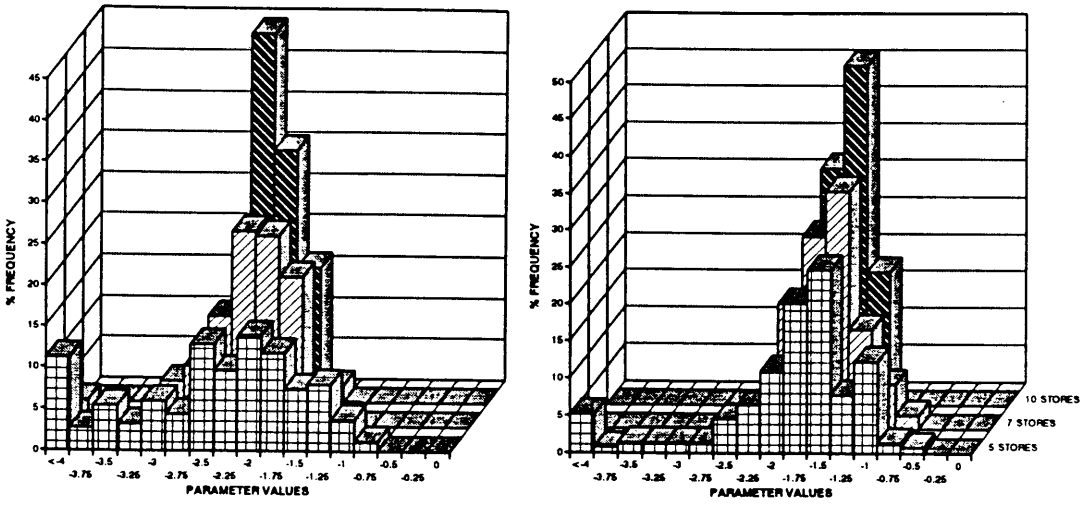

Figures 1a, b: Distance and Chain Image with 14 and 13 store choice sets.

The chain image variable is very consistent across models, and the sensitivity analysis demonstrates the robustness of this variable with changing choice set and choice set size as in figure 1b, providing evidence of the importance of chain image in store choice and how the importance of chain

image varies across market segments. For all three income groups, the chain image parameter indicates a significant preference for the upmarket chains, and this preference tends to increase with income. Likewise, the white market segment also shows a significant preference for upmarket chains. In sharp constrast, the black consumers are less likely to shop at a store belonging to an upmarket chain. Taking the exponential of the parameter estimates from the thirteen store calibrations in figure 1a, we find that low income consumers are generally two times as likely to patronize a store belonging to an upmarket chain, *ceteris paribus*, whereas high income consumers are about five times as likely to select an upmarket store than a downmarket store. The striking difference between the black and white consumer calibrations is clearly evident when one considers that the white consumers are about three times as likely to patronize an upmarket chain store, whilst black consumers are somewhat deterred from patronizing such a store, *ceteris paribus*. In figures 3a and 3b, the distributions of the chain image parameters for the white and black market segments for ten, seven and five stores are presented. The estimates for the white consumers are significant, stable and symmetrical about the baseline estimate, but become increasingly dispersed with smaller choice set size. For the black consumers, the parameters are far less stable with respect to the baseline estimate, and for the five store choice set, are extremely dispersed with most of the parameters more negative than the baseline estimate. The small sample size (48) of black consumers may account for the instability of the parameter estimates in our sensitivity analysis. The $\rho^2$ values throughout the analysis indicate strong model

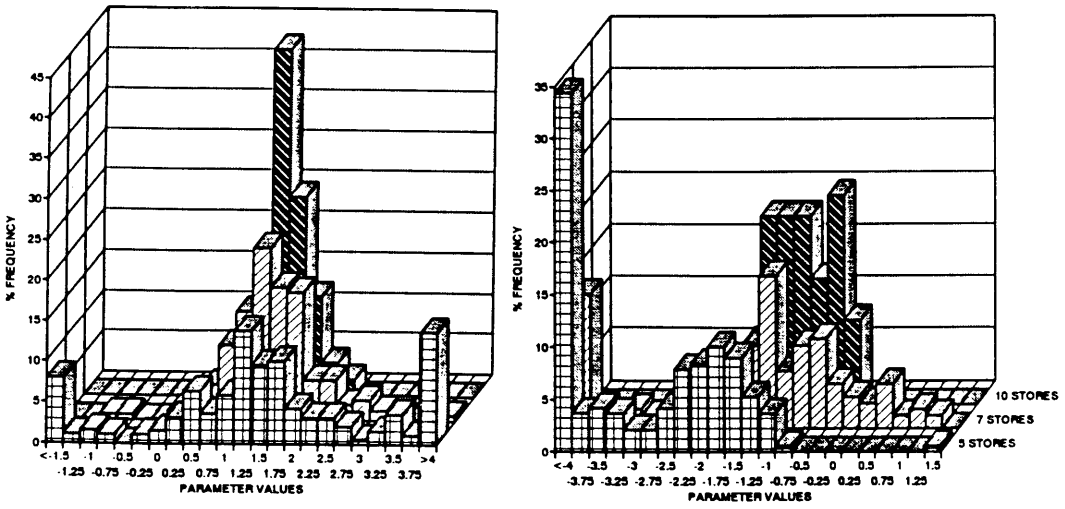
performance for the majority of random choice sets.

(a) (b)




Figures 2a, b: Distance parameter frequency distributions, (a) low, (b) high income groups.

(a) (b)




Figures 3a, b: Chain Image parameter frequency distributions, (a) white, (b) black consumers.

## SUMMARY

This paper presented results from a sensitivity analysis of model parameter estimates to choice set specification in destination choice models. By constraining the choice set in a systematic manner, and calibrating the choice set combinations randomly, the robustness or instability of various model parameters is examined under general conditions. The results indicate parameter sensitivity to choice set size and definition for several core variables (store size, competition), as well as the importance of accounting for the fact that the attributes of the store choice process may vary across different market segments. In particular, strong evidence for the stability of distance deterrence across choice sets, and race and income groups, is provided. It appears that low income consumers are more likely to search for lower prices and be less constrained by distance than their high income counterparts, *ceteris paribus*. Chain image proves to be an important and stable variable in store choice, and the likelihood of an individual selecting an upmarket store increases with income whilst racially segmented consumers show sharp constrasts in their preference for upmarket stores. Also, some level of economic prejudice is indicated for patronage behavior in Gainesville on the part of high income consumers. In general, this analysis serves as a warning to analysts of the dangers of producing misleading results by mis-specifying choice sets, and not segmenting their consumer samples by market segments.

## REFERENCES

Ben-Akiva ... and Lerman S R, 1974, "Some estimation results of a simultaneous model of auto ownership and mode choice to work" *Transportation* **4** 357 - 376

Fotheringham A S, 1988, "Consumer store choice and choice set definition" *Marketing Science* **7** 299 - 310

Fotheringham A S, 1986, "Modelling hierarchical destination choice" *Environment and Planning A* **18** 401 - 418

Fotheringham A S, 1983, "A new set of spatial interaction models: the theory of competing destinations" *Environment and Planning A* **15** 15 - 36

Fotheringham A S and Trew S, 1993, "Chain image and store choice modelling: the effects of income and race" *Environment and Planning A* **25** 179 - 96

Fotheringham A S and O'Kelly M E, 1989, *Spatial Interaction Models: Formulations and Applications* (Kluwer, Dordrecht)

Greene W H, 1991, *LIMDEP: Version 6.0* Econometric Software Inc., New York

Manski C F, 1977, "The structure of random utility models" *Theory and Decision* **8** 229 - 54

McFadden D, 1978, "Modelling the choice of residential location" in *Spatial interaction theory and planning models* Eds A Karlqvist, L Lundqvist, F Snickars, and J W Weibull (Amsterdam, North-Holland) pp 75 - 96

Pellegrini P A and Fotheringham A S, 1994, "Defining spatial choice sets: consumer shopping behavior" Working paper available from the authors

Pellegrini P A, Fotheringham A S, and Lin G, 1994 "Parameter sensitivity to choice set definition in shopping destination choice models" Working paper available from the authors

Stevens S S, 1957, "On the psychophysical law" *Psychological Review* **64** 153 - 181

Thill J-C, 1992, "Choice set formation for destination choice modelling" *Progress in Human Geography* **16** 361 - 382

Williams H C W L, and Ortuzar J D, 1982, "Behavioural theories of dispersion and the mis-specification of travel demand models" *Transportation Research B* **16B** 167 - 219