Abstract

Residential location search has been an interesting topic to both practitioners and researchers. The housing search process starts with an alternative formation and screening practice. At this level households evaluate all potential alternatives based on their lifestyle, preferences, and utilities to form a manageable choice set with limited number of plausible alternatives. Then the final residential location is selected among these selected alternatives. This bilateral decision making process can be used for both aggregate resolution zone selection as well as searching the housing market for potential alternative dwelling types. This paper studies a zonal level household housing search behavior. Initially, a household specific choice set is recognized from the entire possible alternatives in the area based on the average household work distance to each alternative. Following the choice set formation step, a discrete choice model is utilized in this study for modeling the household final dwelling zone selection behavior. A hazard-based model is used for choice set formation module while the final choice selection is modeled using a mixed logit formulation with a sample correction factor and deterministic inter-alternative correlation effects. The approach presented in the paper provides a remedy for the large choice set problem typically faced in housing search models.

Introduction

Residential location search has been an interesting research topic in many fields including transportation, urban planning, geography, economics, and other related disciplines. Metropolitan planning organizations, real estate companies, insurance companies and financial institutions are also among the non-academic organizations that are interested in having an accurate housing search model. Since the early introduction of the discrete choice paradigm, the individual's alternative selection behavior has been primarily modeled using this approach (McFadden 1974). The prediction potential and the accuracy of a discrete choice model parameter estimation process itself are highly reliant on the choice set composition (Ben-Akiva and Lerman 1985 and Timmermans and Golledge 1990). Even though recent advances in computational power allows researchers to work with large datasets, in practical applications, the difficulty of handling many alternatives makes it necessary to reduce the number of alternatives in the choice set into some manageable size. In the literature, there have been two extreme approaches for selecting the set of alternatives; first, randomly selecting a finite number of alternatives from the *universal choice set*, as it is defined by Ben-Akiva and Lerman (1985), second, considering all plausible alternatives (Salomon and Ben-Akiva 1983 and Thill and Horowitz 1991). It can be shown that both approaches can raise some concerns. Although inclusion of all possible alternatives may seem to be a conservative approach, nonetheless, it can be unrealistic as it assumes decision makers have perfect knowledge about all alternatives. This approach can result in assigning non-negative selection probabilities to some alternatives that otherwise may not be known or be available to the decision maker. On the other hand, random selection of few alternatives for the choice set by stratified sampling or other similar approaches can result in bias and possibly inaccurate parameter estimation.

In addition to the two abovementioned approaches, there are other methods to address the choice set formation issue. Disaggregate alternatives can be combined into aggregated sets and form a broad resolution subset of alternatives which consequently result in choice set size reduction. This alternative aggregation method is satisfactorily studied in the literature from different perspectives (Kitamura *et. al* 1979 and Ben-Akiva and Lerman 1985). Alternatively, a selected set of all possible alternatives can be chosen to form a smaller and more manageable choice set using a heuristic or non-heuristic approach, in which alternatives are evaluated by certain criteria for being included in the choice set. The later choice set size reduction method has not been sufficiently studied and it is the main target of this paper. For instance, Arnold *et al.* (1983) utilized a preference ranking method to rank the shopping destination where only stores ranked as *excellent* are included in the individual choice set.

This study aims to introduce a behavioral method for housing search choice set formation followed by an application of this behavioral choice set formation in a discrete choice model. The residential location choice process of this study starts with an alternative evaluation and screening practice. The alternatives are filtered based on average household work distance using the individuals' priorities, lifestyle, preferences, and utilities. While there are several factors

affecting the selection of housing alternatives (e.g., property value, commute distance, school quality, safety, tax rate, etc), in order to show the practicality of the approach, only average work distance is considered for the choice set formation purpose. Nonetheless, land value is also included as an explanatory variable in the final alternative selection model. Having individuals' choice sets in hand among which the final choice will be selected, the final residential location selection behavior is modeled using a mixed logit formulation in which sampling correction factors are included to lessen the unwilling bias affecting the parameter estimations. Additionally the correlation among the alternatives is included in the model using an additional deterministic utility term added to the original utility function of the discrete choice model.

The rest of the paper is organized as follows. First, a brief literature review is presented and the study approach is discussed. Then the choice set design algorithm along with the data used for its model development are explained. Following the choice set formation algorithm discussion, the discrete choice model, data, methodology and results, are presented. Finally, conclusions and future research directions are discussed in the final section.

1. Background and Study Approach

The choice set configuration problem can be traced back to early applications of discrete choice models. Ben-Akiva and Lerman (1985) proposed the stratified sampling procedure to generate the alternative set and showed the efficiency of that approach. Srinivisan (1987) introduced three levels in screening the alternatives and finding the final choice set: awareness set, evoked set and choice set. He borrowed the term evoked set from another study by Howard who originally introduced it in 1963. According to his model, the awareness set consists of all alternatives the consumer is aware of. This set is then filtered to the evoked set which is a subset of the awareness set and consists of those alternatives that meet certain criteria for further consideration. Finally, the choice set is a subset of the evoked set in which there are very few alternatives including the final choice which is the immediate group of alternatives before making a decision. Shocker *et al*. (1991) employed the term consideration set for evoked set which was originally introduced in a study by Wright and Barbour in 1977.

Other than the different definitions for the choice set, various solutions have been introduced to deal with the choice set problem. Willumsen and Ortuzar (2001) listed three ways for tackling the choice set problem available in the literature:

1- Rule-based heuristic or deterministic choice set generation methods,
2- Simply asking the individuals in the survey regarding their preferences about the feasible alternatives,
3- Application of random choice sets.

Regardless of the way that the choice set is designed, if a non-random choice set is formed, the impact of the systematic choice set formation on the successive model estimation (in this study a logit model) should be accounted for. Heckman (1979) introduced a consistent estimator to correct for sample selection bias due to endogenous binary explanatory variables in

linear regression models. In the context of residential location search, a particular type of neighborhood, such as urban/suburban or bike-friendly/bike-non-friendly, is selected using a binary choice equation. Then this latent index equation is endogenously joined to the second regression model. The correlation between the stochastic term of the latent index equation and the regression model indicates the presence of self-selection. Despite the restriction of the Heckman correction method to a binary latent index equation as well as its limitation to regression models, it has been widely cited and used. A comprehensive review of applications of Heckman correction method in criminology literature can be found in Bushway *et al* (2007). Zhou and Kockelman (2008) treated the residential location as a binary (urban/suburban) variable and modeled total household vehicle miles traveled as a continuous variable. Heckman correction method applications are limited to the binary selection cases. Therefore, the successive model which is usually a regression model is conditional and bounded on the selection of a binary selection model.

Multidimensional choice models such as nested logit models obviate the binary limitation of the selection part of Heckman correction method in which the self-selection bias is captured by including a latent utility value in the higher level models (Ben-Akiva and Lerman 1985). However, some other disadvantages are tied with them in the residential location search application. Initially, the alternatives across decision makers are identical. In other words, individuals cannot have alternatives from different nests; instead the lower level aggregate nests are pre-defined across decision makers. Secondly, computationally, total number of alternatives considered for each individual cannot be very large while in a residential locations search the number of alternatives is usually huge. In other words, difficulty of estimation increases as the number of choice dimensions increases (Wen and Koppelman 2001). Consequently in practice, nested logit models with multiple nests are estimated sequentially because simultaneous estimation can be cumbersome.

Although, the above-mentioned binary and multidimensional self-selection approaches are capable of controlling the sampling bias, however they are not behavioral approaches in the case of residential location search. A house searcher does not search all alternatives nor a specific aggregate pre-defined category of the alternatives, those that in reality has not been defined to him/her. In lieu, he/she may employ different search strategies, such as learning-based search and area-based search (Huff 1986) to make a manageable choice set from which the final residence will be selected. Therefore, a compound model composed of a behavioral choice set formation and a discrete choice model can be a suitable candidate representing how decision makers behave in reality (Habib and Miller 2007). Still, in such behavioral approaches the way that sampling bias problem is addressed can be similar to the Heckman correction method and nested logit models, in which a component representing the correlation between the lower and higher level models, is included in the successive model which is a discrete choice model in this case.

Estimation of choice models with sample of alternatives is a well-developed area to which many researchers contributed. If the probability of selecting an alternative in a choice set

is known, the model sampling bias can be alleviated by using that probability. Ben-Akiva and Leman (1985) expansively reviewed the methods for sampling of alternatives and the related techniques for calibrating a logit model based on a designed choice set. It is discussed in their book that the basic logit model can be modified by utilizing an additive alternative-specific correction for the bias. Kanaroglou and Ferguson (1996) generalized the aggregated spatial choice method presented by Ben-Akiva and Lerman in the context of inter-regional migration. Waddell (2000) also employed the introduced correction method of Ben-Akiva and Lerman in developing the residential location and housing market component of UrbanSIM. There are many other applications for the sampling of alternatives in discrete choice analysis (See: McFadden 1978, Ben-Akiva and Watanatada1981). Likewise, this correction method is utilized for adjusting the bias of sampling in this study.

This study introduces an application of discrete choice models with a sample of alternatives in which an innovative behavioral search process for sample design is embedded. Commute distance as one of the most influential factors on residential location is used for the choice set formation (Clark *et al.* 2003 and Waddell 1996). More specifically probability of selecting a residential location area is defined based on its distance to the work location of the household employed members. Then the choice is semi-randomly selected based on these probabilities. Therefore, for each household, a household-specific choice set is formed among which the more desired area is selected based on the utility that the area offers to the household. Average land value of the residential compartments as another important variable in housing search behavior is included among the explanatory variables used in the model development of this study. More detailed discussion about the modeling practice of this study will be presented in the next following sections.

2. Data

Puget Sound Transportation Panel (PSTP) was used as the primary source of data that is used in this study. The PSTP is a panel data for Seattle Metropolitan Area (Murakami and Watterson 1992) .Nonetheless, only household observations of the King and Kitsap county areas are used for the modeling practice due to need for auxiliary data (e.g., property values, etc) that were not available for other two other counties (Snohomish and Pierce counties). The last eight waves out of the existing ten waves in the PSTP covering the last decade of the 20th century plus the two first years of the twenty-first century are included in this study. The PSTP provides a wide range of variables in the household level including household socio-demographic attributes. Furthermore, person level attributes such as home to work distances are also provided in the PSTP.

Average household work distance as the variable that is used for screening the household choice is directly obtained by running some queries on PSTP data. The property value as a critical explanatory variable in the main discrete choice model is not provided in the PSTP. Land values and house prices are mainly attained by county assessment departments. This information

is mainly provided for property tax preparation purposes and is cast away after a decade or so. Such data is available only for King and Kitsap counties at the TAZ and tract levels. Therefore, the PSTP data was filtered based on the counties in which household resides as well.

The data retrieved from the two counties assessment department (King County Assessment Department 2009 and Kitsap County Public Data 2010) is at the very detailed parcel level and it should be aggregated into the census tract level to be coordinated with the PSTP data. Parcel level addresses are mapped to the census street 2000 file and then aggregated up to the tract level as a GIS application. Finally, the aggregated land values and housing prices are merged to the PSTP data.

The built-environment characteristics are borrowed from an adjunct survey of the PSTP in which different job category counts, intersection density, transit availability and many other land-use related variables in a grid of 750 meters by 750 meters are presented.

Finally, historical macroeconomic data are also merged to the abovementioned data sets. Variables like interest rate, inflation rate, gas price and unemployment rate are all tested in the models and their impact on the household decision on residential location attributes are examined.

3. Sampling of Alternatives: Model Formulation and Methodology

As noted earlier, location selection process can be broken into two consequent and correlated sub-processes; initially, household members form their choice sets by screening available alternatives and filtering them based on their priorities, and preferences. Following this step, they single out the most desirable alternative among the filtered alternatives of the choice set. In this section the choice set formation process is discussed in more details.

An extensive curve fitting exercise was undertaken to find the best distribution representing the critical average work distance at which household finally resides. By average work distance, the average work distance of all employed members of a household to their (potential) residential location is intended. It was found that the average work distance follows a Weibull distribution. Table 1 shows the results of the distribution test on work distance based on Kolmogorov-Smirnov statistics (Chakravarti *et al.* 1967 and Eadie *et al.* 1971).

Table 1 Best Fitted Distribution to the Dependent Variables

	Average Work Distance	
Distribution	Kolmogorov-Smirnov Statistic	Rank
Beta	0.04001	2
Chi-Squared	0.13611	15
Exponential	0.15026	16
Gamma	0.04053	3
Gen. Extreme Value	0.04818	5
Laplace	0.15608	17
Log-Logistic	0.09731	11
Logistic	0.11387	13
Lognormal	0.08385	10
Normal	0.10011	12
Weibull	**0.03533**	**1**

** The smaller the KS statistic is for a distribution, the closer that distribution is to the data*

It is assumed that depending on household's attributes, decision makers have some value in mind for the maximum commute distance beyond which housing alternatives will not be attractive to the household. In such cases, thoughts of increasing work distance do not survive and the household will reject any alternative with the distance surpassing the threshold defined for the household. This interpretation of the two continuous dependent variables can suggest using a hazard-based formulation framework. In a mathematical language, this can be formulated as:

$$\lambda(t)dt = \Pr(t + \Delta t \geq T \geq t \mid T \geq t) = \frac{f(t)dt}{S(t)} = \frac{S'(t)dt}{S(t)} \qquad [1]$$

where $\lambda(t)$ is the probability of failure for individual *i* given that it has survived until time *T*, $f(t)$ is failure probability density function and $S(t)$ is the survival function.

The survival function can be calculated using Equation [1] as:

$$S\ (t) = \exp\left[-\int_0^t \lambda(u)du \right] \qquad [2]$$

In addition to the baseline hazard function, other covariates like socio-demographic attributes, built-environment variables and macroeconomic factors can also be incorporated in the hazard function using a proportional hazard formulation which was initially introduced by Cox (1959). The proportional hazard formulation for average work distance with Weibull distribution is as follows:

$$\lambda_i(wd) = \gamma\ wd^{\gamma-1} \exp(-\theta_x X_i) \quad [3]$$

where γ is the shape parameter of the Weibull distribution, X denotes explanatory variables, θ_x is the vector of parameters, and *wd* stands for the average work distance.

Using the same definitions, the survival function with Weibull assumption for the baseline hazard can be shown as:

$$S_i(wd) = e^{-wd^{\gamma} \exp(-\theta_x X_i)} \quad [4]$$

In a mathematical language, the likelihood of failure in accepting a work distance while examining different alternatives is equal to the hazard of failure to accept the alternative times the probability of surviving without accepting it. The likelihood function that is formulated for the average work distance and property value based on their hazard and survival functions across all alternatives, prices, and distances can be written as:

$$L = \prod_{i=1}^{N} \lambda_i(t) \times S_i(t) \quad [5]$$

where N is the number of observations. This function can be maximized to estimate its parameters. The probability density functions estimated by using the results of parameter estimation of Equation [5] are then utilized to generate individual choices.

3.1 Explanatory Variables

The PSTP data set provides a long list of household socio-demographic attributes including income, auto ownership, number of adults, number of workers, among others. Several other dummy variables were generated that represent changes in household status such as lifestyle transitions but were not found to be statistically significant in the model.

Furthermore, one built environment variable and land-use characteristics of the area in which household resides, is included in the models. Frequency of the transit service during the day, especially mid-day, was found significant in the work distance model.

In addition, macroeconomic related factors like inflation rate and unemployment rate were included in the explanatory variable pool. In order to have all prices and income values to be comparable, the first used wave of the PSTP was assumed to be the base year and incomes referring to years after the base year were deflated to the base year using the historical inflation rates. Macroeconomic effects on the household work distance are captured through the unemployment rates obtained from the US 125 years Bureau of Labor Statistics (U.S. Bureau of Labor Statistics 2009).

The average value and standard deviation of the explanatory variables that were found statistically significant in the models are presented in Table 2.

Table 2 Explanatory Variable Used in the Models

Explanatory Variable	Average	St. Dev.
Income	51537.12	26985.79
Number of employed	1.20	0.85
Number of Vehicles	1.76	0.83
Change in Number of adults	0.00	0.45
Mid-day transit availability*	5.09	9.88
Unemployment rate	5.82	1.09

**750 meters by 750 meters gridcells*

3.2 Modeling Results and Analysis

The results of parameter estimation of choice set formation model for average work distance are presented in Table 3. Model parameters are estimated by maximizing the likelihood function presented in Equation [5] using the *nlp* procedure provided by SAS 9.1.3 package. Before evaluating the quality of the estimated parameters, it should be noted that the effect of covariates in a hazard model is facilitated by incorporating negative sign for parameters in formulation. In other words, if a covariate gets a negative sign, the chance of failure or the probability of accepting a work distance is increased. Alternatively, having a positive sign means that any increase in the covariate decreases the chance of failure for the household which implies that the household tends to increase the work distance.

Table 3 Results of Joint Model of Household Average Work Distance

	Household Average Work Distance		
Parameter	Estimate	t Value	Pr > |t|
Sigma	1.828	12.505	0.000
Constant	2.847	6.680	0.000
Previous Work Ditsance	0.074	4.274	0.000
Change in Income (X100,000)	0.683	1.165	0.246
Number of Vehicles	0.281	2.195	0.030
Number of Employeds	0.193	1.834	0.069
Change in Number of Adults	-4.630	-1.740	0.084
Mid-Day Transit Availability	-2.398	-3.576	0.000
Unemployment Rate Change	-0.238	-1.486	0.139
Likelihood value with only constant		*-439.39*	
Likelihood value at convergence		*-410.55*	
$-2[L(C)-L(\beta)]$		*57.68*	

The Weibull distribution of the work distance models has a monotonically increasing shape because sigma parameter is greater than one.

It was found that household's current average work distance is considerably affecting the household decision about its new residence. Annual income which is positively correlated with the number of vehicles is also important on the household decision about the average work distance. The higher a household income is, the farther they can select their work location from the residence. Wealthier households are also more likely to live in suburban areas and commute farther distances. Similarly, total number of vehicles in the household is positively correlated with the work distance. Households with more workers can commute to farther work destinations whereas households with more changes in the number of adults are likely to work closer to their home. Households living in areas with more available mid-day transit are also more likely to reduce their work distance. Finally, unemployment rate as a representative of the supply side of the market, found to be significant in the household average work distance. Results shown in Table 3 imply that any increase in the unemployment rate shifts the households' tendency to reduce their average work distance.

The likelihood function value at convergence is -410.55 Therefore, the statistic $-2[L(C)-L(\beta)]$ would be $-2[410.55-439.39] = 57.68$. It is noteworthy that this statistics is asymptotically Chi-square distributed with degrees of freedom of 10 which is highly significant.

3.3 Simulation and Sensitivity Analysis

The developed housing search choice set formation model was evaluated evaluating by its overall goodness of fit, likelihood value at convergence, and estimated parameters that were all statistically significant. The explanatory variables were also selected such that choice specific, household taste variation and market characteristics are included in model.

In this section its performance through a simulation practice is examined. The parameter estimates of the model that are presented in Table 3 were used to estimate the probability of accepting a work distance for each household. As noted earlier, the probability density function for accepting a work distance can be obtained by estimating the product of hazard and survival functions. The probability density function can be easily written using Equations [2] to [4] as:

The probability density function for work distance is:

$$f_i(wd) = \left[\gamma\ wd^{\ \gamma\ -1} \exp(-\theta_x X_i)\right] \times \left[e^{-wd^{\gamma} \exp(-\theta_x X_i)}\right] \qquad [6]$$

As shown in Equation [6] above, the probability density functions of work distance is a function of household characteristics. The probability of accepting a work distance is estimated for each household using Equation [6]. This equation can generate a probability density function for each household similar to the one shown in Figure 1 for average household work distance.

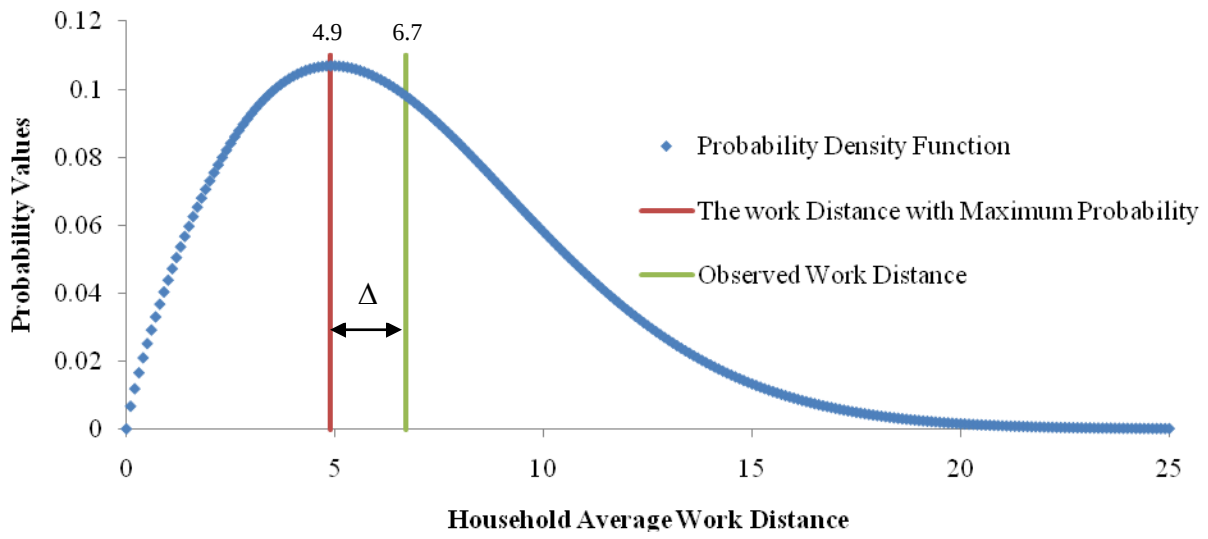

Figure 1 An example of estimating the percentage distance between the maximum probability density and the observed value of household work distance ($\lambda = (6.7 - 4.9)/4.9 = 36.73\%$)

Given the observed household work distance, one can compare and estimate the difference (Δ) between the observed and estimated maximum probability values for each household. Then the percentage difference (λ) between the observed and simulated prices and work distance values can be calculated for each observation.

Various distributional forms are fitted to λ values and the best fitted density functions are selected based on the Kolmogorov-Smirnov test (Chakravarti *et al.* 1967 and Eadie *et al.* 1971) that was shown to be statistically highly significant in both cases. The cumulative density function of the best fitted distribution for work distance variables is presented in Figure 2.

Figure 2 presents the cumulative density functions for the estimated percentage differences between the observed and simulated work distance along with the best fitted density function parameter values.

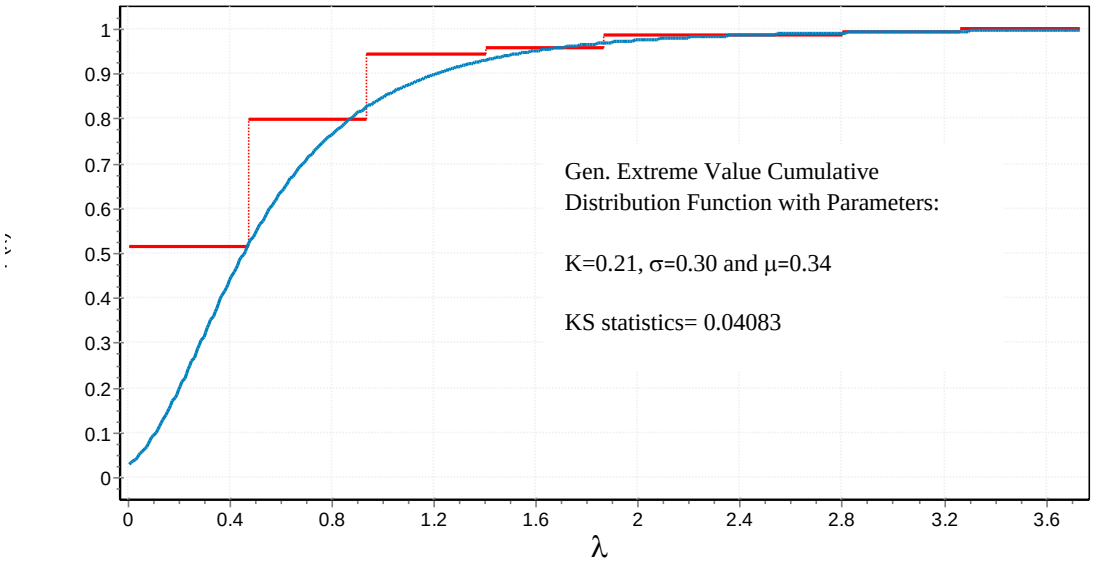

Figure 2 Sample frequencies and cumulative density function for percentage differences between simulated and observed average work distance.

As shown in Figure 2, over 80 percent of the predictions are at most 80 percent different from the observed work distance. Other than model validation purpose, Figure 2 can indicate that the model developed in this study are capable of being used in a simulation process in which household housing search choice sets can be accurately generated. Therefore, there is no need to consider a *universal choice set* for all the households which will cause computational difficulties and potential risk of erroneous estimates. Rather, the choice set can be cut down into a manageable but highly accurate set of alternatives that can result in more efficient and unbiased model estimation. This task will be explained in the next section.

4 Sample of Alternatives Generation

Out of the 824 Transportation Analysis Zones (TAZ) in the Seattle Metropolitan Area, 741 of them are included in the *universal choice set* available to the households living in the area among which they select their residences.

Using the probability density function of Equation [6], for each household, the most probable work distance is simulated around which the probability of residing is the highest (*desired work distance*). Other than the desired work distance for each household, the average distance to the household employed members work locations is calculated for each one of the entire 741 zones in the area. Therefore, for each household, 741 figures are calculated representing how far on average the work distances of household members will be if the household moves to a zone in the area (*actual work distance*). Having these two distances in hand (*actual work distance* and *desired work distance*) for all household, probability of moving to any of the 741 zones across all the households in the data is defined. This probability is later used for sampling the alternatives (741 zones) into a smaller set of choices. This probability is

estimated as the exponential of the normalized (by *desired work distance*) difference between the *desired work distance* and the *actual work distance* if the *desired work distance* is smaller than the *actual work distance* while if this is not the case, exponential of negative normalized (by *desired work distance*) *actual work distance* represents the probability of selecting that zone. Based on the way that the probabilities are constructed, it is intuitive that, probability of selecting a zone increases as it gets closer to job locations of household members while it decreases when households considers farther zones beyond the *desired work distance*.

A subset of all alternatives (zones) is randomly selected based on the estimated probabilities for each household. This pseudo-random selection process starts with determining a value for total number of draws and ends with providing a list of alternatives for each household. Number of alternatives selected for each household is not made fixed so that it would be compatible with what happens in reality. Alternatives are selected with replacement and the alternatives with higher probabilities have a greater chance to be selected. Nonetheless, each alternative is included in the choice set only once and if it is selected more than once that random draw is void.

In order to approximate the most appropriate choice set size, nine total random draw values are examined. Table 4 shows the effectiveness of these random draw value scenarios.

Table 4 Evaluating the effectiveness of different random draw values

Random Draws	Truly Included Final Decision (1)	Average Choice Set Size (2)	(1)/693 (%)	(2)/741 (%)
25	94	23	13.56%	3.10%
50	167	43	24.10%	5.80%
100	241	77	34.78%	10.39%
200	367	128	52.96%	17.27%
300	424	165	61.18%	22.27%
400	446	195	64.36%	26.32%
500	506	219	73.02%	29.55%
600	518	239	74.75%	32.25%
700	524	255	75.61%	34.41%

The first column in Table 4 shows total number of random draws performed for forming the choice set. The second column shows the total number of households whose final residential location decision has been included in the choice set. The third column presents the average choice set size for the household. In total 693 households from King and Kitsap counties are included in this study to which 741 zones were available for choosing their next residential locations. The fourth column is calculated by dividing the second column to 693 which is the total number of households. Therefore, it represents the percentage accuracy of the choice set formation algorithm. Finally, the last column shows the percentage of alternatives that have been included in the final choice sets. There are two important factors in evaluating a choice set generator algorithm: the predictability capability of the algorithm and size of the generated choice sets. Unfortunately, these two factors are negatively correlated; therefore an equilibrium point should be selected by the researcher at which the choice set size is acceptably small while

most of the time the actual decision is included in the choice set. In other words, increasing choice set size raises the chance of not excluding the decision's maker final choice, while it is also counter the willingness of shrinking the choice set size. Behaviorally, people do not compare a large set of alternatives, instead, a small set of most desirable choices are selected among which the final choice is chosen. Although, the final decision is manually included in the choice set for the model development step, for simulation purposes, it is critical to have a choice set formation algorithm that does not exclude the most important alternatives which are usually selected by the decision makers from the choice set.

Figure 3 shows the tradeoff between the accuracy of the presented choice set formation algorithm of this study across different choice set sizes. It can be discerned from Figure 3 that the choice set formation algorithm of this study has an acceptable performance, because if only one third of the *universal alternatives* are selected by this algorithm, then, 75% of the times the final selected choice is not excluded from the choice set.

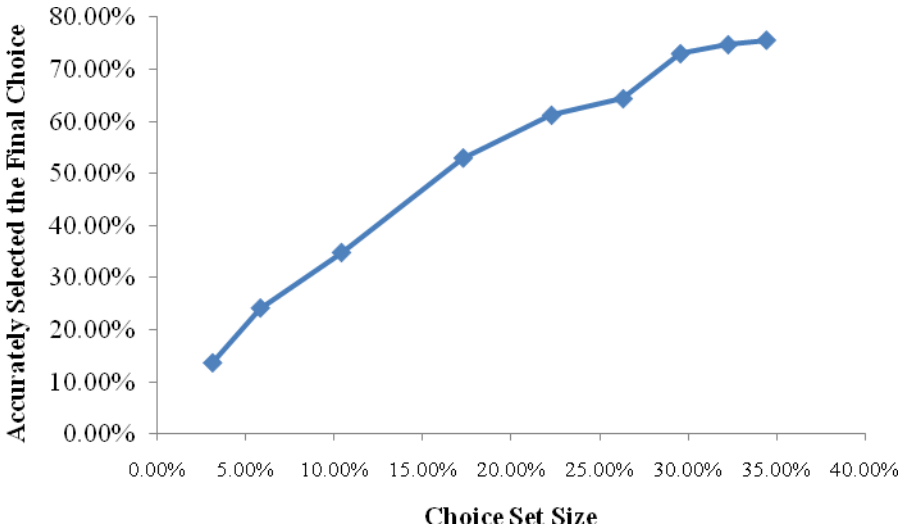

Figure 3 Tradeoff between choice set formation algorithm accuracy and choice set size

In this paper, it is attempted to diminish the effect of specific choice set on the parameter estimation of the discrete choice model by utilizing a latent index in the successive model which will be elaborated in more details in the next section.

5 Residential Location Choice Model Methodology

Different choice set compositions can have singular impacts on a discrete choice model. Therefore, it is very critical to adjust these effects on the parameter estimations. Otherwise the estimated parameters are not consistent anymore. This study utilizes the method presented by Ben-Akiva and Lerman (1985) and also applied in a route choice selection application by Frejinger *et al.* (2009). More specifically, a multinomial logit model is developed on a subset of

the entire alternatives which are selected based on their probabilities of being selected in the choice set. It has been proved (McFadden, 1978) that the multinomial logit model can be consistently estimated on a subset of alternatives using classical conditional maximum likelihood estimation. The probability that an individual *i* chooses an alternative *j* can be formulated as:

$$P_{ij} = \frac{e^{\mu V_{ij} - \ln \frac{q_{ij}}{\sum_{k=1}^{K} q_{ik}}}}{\sum_{l=1}^{L} e^{\mu V_{il} - \ln \frac{q_{il}}{\sum_{k=1}^{K} q_{ik}}}} \qquad [7]$$

where μ is a scale parameter and V_{ij} is the deterministic utility, *K* is total number of alternatives (741) and *L* is the total number of alternatives in the choice subset. The $\ln \frac{q_{ij}}{\sum_{k=1}^{K} q_{ik}}$ alternative specific term corrects for sampling bias. Roughly speaking, q_{ij} represents exponential of subtraction between the most *desired work distance* and the alternative of residential location distance to the household employed members' work locations (*actual work distance*). The book by Ben-Akiva and Lerman (1985) can be referred to for more detailed discussion on sampling of alternatives and further examples on this topic.

Therefore, by using Equation [7], we are assured that multinomial logit model is consistently developed and the sampling bias is corrected meanwhile attractive alternatives with higher probability than unattractive alternatives are included among the chosen set of alternatives.

In order to capture the unobserved heterogeneity across households and their sensitivity to observed exogenous variables a mixed logit model is utilized with a random parameters specification in this study. Briefly saying, the deterministic utility of Equation 7 can be re-written as:

$$V_{ij} = \beta x_{ij} = (\beta^{m} + \beta^{s} \eta) x_{ij} \qquad [8]$$

Where β^{m} and β^{s} are respectively fixed mean and scale parameters, and the stochastic component, η , is assumed to be standard normal. The choice probabilities are estimated using a Monte Carlo simulation method in which Halton quasi-random sequences (Halton 1960) numbers are randomly drawn.

6 Residential Location Choice Model Results

In this section results of the developed mixed logit model are presented. But initially, a descriptive analysis for the explanatory variables used in the modeling practice of this study is

presented. Table 5 shows the average and standard deviation for the dependent variables that are used in the modeling practice of this study. Considerably more number of explanatory variables were tested in this study while the variables found to be statistically significant in the final model are reported in Table 5.

Table 5 Explanatory variables used in the mixed logit model

Parameter	Name	Average	St. Dev.
Number of real estate and rental jobs*	J_Real	1.43	0.39
Industrial Square Feet*	IndSqFt	2733.16	5516.38
Average AM transit availibility*	AM	8.06	8.63
Number of manufacturing jobs-Neighbors*	J_Manu_N	48.79	30.82
Commercial Square Feet-Neighbors*	ComSqFt_N	505479.56	281753.85
Govermental Square Feet-Neighbors*	GovSqFt_N	86688.90	52077.74
Average AM transit availibility-Neighbors*	AM_N	8.64	4.31
Population*Unemployment Rate/Area**	UEP	10617.08	10673.60
Number of HHlds / Area**	HHld	532.60	558.15
Number of seniors 65-74 / Area**	Seniors	100.60	125.75
Number HHlds with Income >150K /HHlds Income**	Rich	150.46	129.52
Number HHlds with Income between 60K-75K /HHlds Income**	Medium	350.62	124.96
Number HHlds with Income between 15K-18K /HHlds Income**	Low	66.52	52.02
Number of bus riders / Gas Price**	Bus	704.81	492.52
Number of Drive Alones/ Gas Price**	Dr_Alone	60.06	13.87
LandValue / (Income X 1000)**	LandVal	49.47	11.50

** 750 meters by 750 meters gridcells*
*** TAZ*

The explanatory variables used in this part are observed at two geographical resolutions. Land use variables relating to the job type totals in a zone are provided by Puget Sound Regional Council are at the fine resolution of 750 meters by 750 meter gridcells while the rest of variables (except for land value) are borrowed from CTPP data files which are available at the TAZ level. Land values as it was discussed earlier in the data section, are in hand from the assessment department data bank. The land values are reported along with the address of the properties, so they have been aggregated in this study to the TAZ level to be compatible with the other explanatory variables.

The first two variables in Table 5 represent the employment situation of the area while the third variable stands for transit availability in the area to which household may move. The next four explanatory variables are included in the model to enhance the model of considering spatial correlation. These four variables represent the land use conditions in the zones surrounding the one which a household considers for its future residence. Unemployment rate of the years during which the relocations occurred are used for unemployment population calculation, then the unemployed population in each zone is divided by the area of that zone to be comparables among zones. Total number of households and number of seniors in the region are also considered in the final developed models to more specifically explain the characteristics of zones. Populations of three income groups are included among the explanatory variable representing high, medium and low income ranges. These income categories are divided by household annual income to be more

compatible with decision makers' behaviors. In other words, as the ratio of number of rich people of a zone to the income of a household increases, that household becomes less interested in living in such zones. Finally mode choice priorities relatively to gas price were included in the pool of explanatory variables.

After a brief discussion on the definitions of the utilized explanatory variables, detailed results of the developed mixed logit model is presented and discussed. Table 6 shows the estimated parameters of the mixed logit model with 200 random runs for choice set generation (on average choice set size of 128).

Table 6 Mixed logit model development results

Parameters	Estimation	t-value
LandVal (βm)	-0.0143	-3.16
LandVal (βs)	-0.0111	-2.14
Correction Factor	0.1486	5.08
J_Real	-0.2684	-4.24
IndSqFt (βm)	-0.00003	-1.95
IndSqFt (βs)	-0.000029	-1.82
AM	0.0152	3.2
J_Manu_N (βm)	-0.002028	-1.81
J_Manu_N (βs)	0.002673	1.37
ComSqFt_N	3.54E-07	2.05
GovSqFt_N	-6.55E-07	-1.11
AM_N	-0.0305	-2.99
UEP (βm)	-0.00004	-5.9
UEP (βs)	0.0000183	4.68
HHld	0.000378	5.05
Seniors	-0.000394	-1.28
Rich	-0.000611	-1.58
Medium	0.00019	1.74
Low (βm)	0.000649	2.7
Low (βm)	0.000544	1.8
Bus (βm)	-0.00222	-1.73
Bus (βs)	0.002309	1.16
Dr_Alone	0.000909	7.67

Model Goodness-of-Fit		
Likelihood Ratio (R)	395.03	2 * (LogL - LogL0)
Upper Bound of R (U)	6637.7	- 2 * LogL0
Aldrich-Nelson	0.3644	R / (R+N)
Cragg-Uhler 1	0.4364	1 - exp(-R/N)
Cragg-Uhler 2	0.4364	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.4463	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.4165	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.0595	R / U
Veall-Zimmermann	0.4022	(R * (U+N)) / (U * (R+N))

General model goodness-of-fit seems very promising based on the results presented on the right hand side of Table 6. Land value as a very critical variable in selecting the zone to which a household decides to move, found to be a statistically significant variable in the model. The stochastic taste variation term for this variable also found to be important to be included in the model. It can be interpreted from the negative sign of LandVal parameter that zones with greater relative average land values than household income become less attractive to the household since they become less affordable to them. Decision makers are less interested in zones with higher real estate jobs as well as industrial zones. This can be rationalized by the fact that these types of

zones are not necessarily family friendly neighborhoods. Availability of AM peak transit service was found to be encouraging people to move to a zone. It is a promising finding for transit service providers because it says that people are interested in residing in areas where transit is accessible. More interestingly, it was found that availability of transit in neighbor zones of a zone is not interesting to a searcher. This means, people prefer to move to other zones in which transit is accessible instead of a zone in surrounding which transit is accessible. It was shown in Table 6 that utility of moving to a zone is magnified if the zone is surrounded by commercial zones while it is reduced if it is bordered with zones with many governmental offices. The findings of this study confirm the intuitive that zones with higher unemployment rates are less interesting to residential location seekers; instead, family oriented zones with relatively higher number of households are more attractive to searchers. The *Seniors* parameter found to be negative which implies that the utility function shrinks if number of seniors increases in a zone. This variable can indirectly reflect the average income of the zone. Parameter estimations for income related variables release interesting results. If the relative number of rich people in a zone to the income of the household increases, this makes that zone less attractive to house searchers. Oppositely, increase in the *medium* variable increases the attractiveness of a zone. Bus ridership found to be a highly taste sensitive variable in the developed model, because the stochastic part of the estimated parameter of this variable found to be greater than the fix part of it. Therefore, depending on other variables such as gas price, household income, congestion level and many other unobserved factors people may be more/less interested in zones with greater/smaller number of bus riders. It was found that TAZs with more lonely drivers are more attractive to residential location searchers. These zones can be categorized as less dense and more suburban looking areas.

The final analysis conducted in this paper discusses the effectiveness of the employed sample correction method. Although a sample size of average 128 was selected for the final analysis, the modeling results of 43, 77 and 165 (50, 100 and 300 runs) sample size also showed no more than 42% difference on average between the presented results in Table 6 and the estimated parameters for these three models. Even if a complete random sample is drawn for each household (100 runs) the parameter estimations are at least 300% of what is presented in Table 6. Therefore, it can be concluded that the utilized correction factor can auspiciously stabilize the parameter estimations while it provides a way of including behavioral choice sets in the discrete choice model.

7 Conclusions and Future Directions

This study presented a behavioral model of alternative set formation for residential location choice problem as well as its application in mixed logit discrete choice model. Briefly, a two-step approach is considered in which alternatives are evaluated and screened based on household priorities, lifestyle, and preferences and for each alternative, the probability of being selected in the choice set is estimated. Following that, the choice set is randomly formed, and then from the

generated choice set the alternative with highest utility can be selected by using traditional choice models. The sampling bias is adjusted in this study by using the sampling of alternatives methods that can be found in a book by Ben-Akiva and Lerman (1985). An innovative and behavioral sample design method was introduced in this study which uses household average word distance as the yardstick for evaluating the alternatives. A hazard-based formulation with Weibull distribution was employed for modeling development of sample selection process. A choice set was simulated for each decision maker using the developed choice set formation model. Finally the simulated choice sets were used in a mixed logit model to model the disaggregate behavior of decision makers in finding a residential location area.

The Puget Sound Transportation Panel of Seattle Metropolitan Area was used in this study for the modeling practice along with other sources of data such as built environment, land-use, and economic factors.

The models developed in this study were validated in different ways and overall, it was shown that they are capable of generating highly accurate choice sets that can result in more efficient and unbiased housing search models.

Further improvements to the model include: incorporating heterogeneity in the choice set formation, investigating the importance of other variables on housing search choice set formation besides work distance, and including the stochastic correlation between the alternatives in the mixed logit model. These improvements remain as future research tasks. It should be also noted that the application of the proposed modeling framework is not limited to the housing search problem. Such a framework can be used in other contexts where large number of alternatives should be evaluated. For instance, in the case of activity location choice (e.g., shopping) a similar approach can be used, however, instead of price and distance, other appropriate factors such as size (e.g., number of stores, or retail jobs) can be used along with distance.

8 Acknowledgement

The authors are grateful to the Puget Sound Regional Council and Mr. Neil Kilgren for providing the Puget Sound Transportation Panel data. Partial support of this research was provided by the 2009 Freeman Fellowship of the American Society of Civil Engineers.

9 References

Arnold, S. J., Oum T. H. and Tigert D. J., (1983), Determinant attributes in retail patronage: seasonal, temporal, regional, and international comparisons, *Journal of Marketing Research*, 20, pp. 149-57

Ben-Akiva M. and Lerman S.R., 1985, Discrete choice analysis. Theory and application to travel demand, Cambridge: MIT Press

Ben-Akiva M. and T. Watanatada, (1981), Application of continuous choice logit model, in structural analysis of discrete data with econometric applications, C. Manski and D. McFadden Eds. MIT Press, Cambridge Mass.

Bushway S. D., B. D. Johnson, and L. A. Slocum, (2007) Is the magic still there? The relevance of the Heckman two-step correction for selection bias in criminology, *Journal of Quantitative Criminology*, 23:151–78

Chakravarti I.M., R.G. Laha, J. Roy, 1967, Handbook of Methods of Applied Statistics, vol. I, John Wiley and Sons

Clark W. A. V., Huang Y. and Withers S. D., (2003), Does commuting distance matter? Commuting tolerance and residential change, *Regional Science and Urban Economics* 33, pp. 199-221

Cox, D. R., (1959), The analysis of exponentially distributed life-time with two types of failures, *Journal of Royal Statistical Society*, Vol. 21B, pp. 411-421

Eadie, W.T., Drijard D., James F.E., Roos M. and Sadoulet B., 1971, Statistical Methods in Experimental Physics. Amsterdam: North-Holland, 269-271.

Frejinger E., M. Bierlaire and M. Ben-Akiva, (2009), Sampling of alternatives for route choice modeling, Transportation Research Part B, Vol.43, 10, pp. 984-994

Habib M. A. and E. J. Miller, (2007), Modeling residential and spatial search behavior: evidence from the Greater Toronto Area, Sixth Triennial Symposium on Transportation Analysis Phuket Island, Thailand, 10-15 June 2007

Halton J.H., (1960), On the efficiency of certain quasi-random sequences of points in evaluating multidimensional integrals, *Numerische Math*, 2, 84-90

Heckman J. J., (1979), Sample Selection Bias as a Specification Error, *Econometrica*, 47(1), pp. 153-161.

Howard J.A., (1963), Marketing Management, Homewood, IL, Richard Irwin

Huff J.O., (1986), Geographic regularities in residential search behavior, *Annals of the Association of American Geographers*, 76, pp. 208-227

Kanaroglou P. S. and M. R. Ferguson, (1996), Discrete spatial choice models for aggregate destinations, *Journal of Regional Science*, 36 (2), pp. 271-90

King County Assessment Department, (2009), Parcel Level Property Values, King County, Washington, http://info.kingcounty.gov/assessor/DataDownload/default.aspx last accessed on May 2010

Kitsap County Public Data (2010) http://kcwppub3.co.kitsap.wa.us/pub_disc/ last accessed on May 2010

Kitamura R., L. Kostynuil and K. L. Ting, 1979, Aggregation in Spatial Choice Modeling, *Transportation Science*, 13, pp. 325-342

Lerman S.R., (1985), Random utility models of spatial choice. In Hutchinson, B.G., Nijkamp, P., Batty, M., editors, *Optimization and discrete choice in urban systems*, Berlin: Springer-Verlag, pp. 200-217

McFadden D., (1974), Conditional Logit Analysis on the Temporal Stability of Disaggregate Travel Demand Models, Transportation Research Part B, 16, pp. 263-278

McFadden D., (1978), Modeling the choice of residential location, in spatial interaction theory and residential location, (1978), *A. Karlquist et al., eds.* North Holland, Amsterdam, pp. 75-96

Murakami E. and Watterson W. T., (1992), The Puget Sound transportation panel after two waves, *Transportation*, Vol. 19, No. 2, pp. 141-158

Salomon I. and Ben-Akiva M., (1983), The use of the life-cycle concept in travel demand models, *Environment and Planning A*, 15, 623-38

Shocker A. D., Ben-Akiva M. E., Boccara B. and Nedugadi P., (1991),Consideration set influences on consumer decision-making and choice: issues, models and suggestions, *Marketing Letters*, 2, pp. 181-197

Thill J. C. and Horowitz J. L., (1991), Estimating a destination-choice model from a choice-based sample with limited information, *Geographical Analysis*, 23, pp. 298-315

Timmermans, H.J.P. and Golledge, R.G., 1990, Applications of behavioral research on spatial problems II: preference and choice, Progress in Human, Geography 14, pp. 311-54

U.S. Bureau of Labor Statistics (2009), Local Area Unemployment Statistics, http://www.bls.gov/lau last accessed on July 2009

Waddell P., (2000), A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim, *Environment and Planning B*, 27, pp.247-263

Wen C. H. and Koppelman F.S., (2001), The generalized nested logit model**,** *Transportation Research B*, 35(7) 627-641.

Willumsen L. G. and Ortuzar J. de D., (2001), Modelling Transport, John Wiley & Sons, New York

Wright P. and Barbour F., (1977), Phased decision strategies. In: M. Starr and M. Zeleny (eds.) management Science, Amsterdam, North Holland, pp. 91-109

Zhou B. and K. M. Kockelman, (2008), Self-selection in home choice: use of treatment effects in evaluating relationship between built environment and travel behavior, *Transportation Research Record*, 2077, pp. 54-61