Transportation Research Part A 46 (2012) 1097–1107

Contents lists available at SciVerse ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection

Taha Hossein Rashidi [a,*], Joshua Auld [b,1], Abolfazl (Kouros) Mohammadian [b,2]

[a] *Department of Civil Engineering, University of Toronto, Sandford Fleming Building, 10 King's College Road, Room 3001, Toronto, Ontario, Canada M5S 3G4*
[b] *Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 W. Taylor St., Chicago, IL 60607, United States*

ARTICLE INFO

Article history:
Received 22 December 2010
Received in revised form 27 January 2012
Accepted 31 January 2012

Keywords:
Housing search
Sampling correction
Choice set formation
Multinomial logit
Hazard models

ABSTRACT

Residential location search has become an important topic to both practitioners and researchers as more detailed and disaggregate land-use and transportation demand models are developed which require information on individual household location decisions. The housing search process starts with an alternative formation and screening stage. At this level households evaluate all potential alternatives based on their lifestyle, preferences, and utilities to form a manageable choice set with a limited number of plausible alternatives. Then the final residential location is selected among these alternatives. This two-stage decision making process can be used for both aggregate zone-level selection as well as searching disaggregate parcel or building-based housing markets for potential dwellings. In this paper a zonal level household housing search model is developed. Initially, a household specific choice set is drawn 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 for modeling the final residential zone selection of the household. A hazard-based model is used for the choice set formation module while the final choice selection is modeled using a multinomial logit formulation with a deterministic sample correction factor. The approach presented in the paper provides a remedy for the large choice set problem typically faced in housing search models.

1. Introduction

Residential location search has been recently become an important 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 working housing search models. Since the early introduction of the discrete choice paradigm, such models of housing selection behavior have often been developed using this approach (McFadden, 1974). As has long been known, the predictive potential and accuracy of a discrete choice model itself are highly reliant on the choice set composition (Ben-Akiva and Lerman, 1985 ; Timmermans and Golledge, 1990). Therefore, in cases of spatial models, as in residential location choice, the handling of the choice set development becomes very important. Even though recent advances in computational power

[*] Corresponding author. Tel.: +1 647 638 5450.
E-mail addresses: taha.hosseinrashidi@utoronto.ca (T.H. Rashidi), auld@uic.edu (J. Auld), kouros@uic.edu (A. Mohammadian).
[1] Tel.: +1 312 996 0962; fax: +1 312 996 2426.
[2] Tel.: +1 312 996 9840; fax: +1 312 996 2426.

0965-8564/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tra.2012.01.007

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 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-zero 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 a few alternatives for the choice set by stratified sampling or other similar approaches can result in biased and possibly inaccurate parameter estimation.

In addition to the two aforementioned approaches, there are other methods to address the choice set formation issue. Disaggregate alternatives can be combined into more aggregated sets 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; 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 method has not been sufficiently studied and it is the main target of this paper (see for example Arnold et al., 1983).

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 step. The alternatives are filtered based on average household work distance using the individuals' priorities, lifestyle, preferences, and utilities. Note that the use of average work distance in choice set composition implies that residential choice is in this instance conditional on workplace choices of the individual's in the household. While there are several other factors that clearly affect 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 in the choice set formation stage. The remaining variables will be accounted for in the final location choice model. The final residential location selection behavior is modeled using a multinomial logit formulation with the sampled choice sets, in which sampling correction factors are included to remove the sampling bias affecting the parameter estimations introduced from the choice set formation stage. Additionally, the systematic spatial dependencies among the alternatives are included in the model using an additional deterministic utility term added to the original utility function of the discrete choice model.

It is important to reiterate that this paper only discusses the household housing search behavior which is conditional on the household residential relocation timing decision. It has been previously discussed elsewhere (Rashidi et al., 2011) that the timing of the residential relocation is endogenously correlated with household employed members' job relocation timing decisions. In the current paper, therefore, the search behavior of the household for finding the most appropriate zone is modeled given that workplace choices are fixed. Note, however, that the converse situation of workplace choice being dependent on household location could similarly modeled in the manner presented in this work with little modification – although it would be necessarily conducted at the individual rather than household level.

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 development are presented. Following this discussion, the discrete choice model, data, methodology and results, are given. Finally, conclusions and future research directions are discussed.

2. Background and study approach

Choice set formation has been the topic of many research studies following the introduction of the discrete choice model method. In this section, an overview of previous work regarding various methods to form an appropriate choice set is presented. This overview starts by discussing simple binary choice set formation methods, continues to discuss nested choice set models and finally more advanced probabilistic approaches used to construct a choice set are reviewed.

The choice set formation 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. Spiggle and Sewall (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. These include using deterministic rule-based frameworks, asking individuals about the alternatives considered during the survey process and using random choice set generation methods. Regardless of the way that the choice set is designed, if a non-random choice set is formed, the impact of the departure from random selection in the 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 the 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 these binary selection cases. Therefore, the successive model which is usually a regression model is conditional on the binary selection.

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 context. 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. Second, computationally, the total number of alternatives considered for each individual cannot be very large while in a residential locations search the number of alternatives is usually extensive. In other words, the 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, they are not behavioral approaches in the case of residential location search. A house searcher does not search all alternatives or a specific aggregate pre-defined category of the alternatives. In actuality they 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 residential location 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.

The estimation of choice models using the sampling of alternatives is a well-developed area to which many researchers have 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 Lerman (1985) 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 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 Watanatada, 1981). 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 choice set formation is embedded. As commute distance is one of the most influential factors on residential location, it is used as the primary covariate for the choice set formation process (Clark et al., 2003 and Waddell, 1996). More specifically, the probability of selecting a residential location area is defined based on its distance to the work locations of the employed members of the household. Then the choice set is semi-randomly selected based on these probabilities. Therefore for each household, a household-specific choice set is formed among which the more desirable areas are selected based on the utility that the areas offers to the household in terms of commuting. The average land value of the residential locations, as another important variable in housing search behavior, is included among the explanatory variables used in the model development. More detailed discussion about the modeling practice of this study will be presented in the following sections.

3. Data

The Puget Sound Transportation Panel (PSTP) was used as the primary source of data used in this study. The PSTP is a set of panel data for the Seattle Metropolitan Area (Murakami and Watterson, 1992) .In this study, only the household observations of the King and Kitsap county areas are used for the model development 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 21 century are included in this study. The PSTP provides a wide range of variables at the household level including household socio-demographic attributes. Furthermore, person level attributes such as home to work distances are also provided in the PSTP. All observations where a household makes a household location change from the previous wave are used in the model estimation, to ensure that only the factors impacting the decision at the time the decision is made are included in the model. Overall this leaves a total of 741 household location decision observations used in developing the model.

The average household work distance is the primary variable that is used for limiting the household location choices and is directly obtained from the PSTP data by averaging all of the commuting distances for working members of the household. It is important to note that this implies that household location choice is conditional on known commute distances, i.e. that the work locations are fixed. It is clear from past research and from direct observation of the current data set, however, that this is not strictly the case in actuality. There is, in fact, some endogeneity in this process, with residential moves occasionally motivating work location changes as observed in research by Rashidi et al. (2011). However, it should be noted that the cases that residential relocation occurred prior to job relocation are not particularly common with only 15% of household location moves in the dataset followed by a subsequent job move. The remaining observations are either residential moves following job moves, residential moves with no job moves observed, or concurrent residential and job moves (36%). So in 85% of cases the assumption that the commute distance is known when making the residential location decision is not problematic.

The property value as a critical explanatory variable in the main discrete choice model, however, is not provided in the PSTP. Land values and house prices are mainly obtained by county assessment departments. Such data were only available from King and Kitsap counties at the TAZ and tract levels for this study. The data retrieved from the two counties (King County Assessment Department, 2009 and Kitsap County Public Data, 2010) were at the very detailed parcel level and were aggregated into the census tract level and coordinated with the PSTP data using a GIS application. The built-environment characteristics were 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 were provided in a grid of 750 m by 750 m. Finally, historical macroeconomic data are also merged into the abovementioned data. Variables such as 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.

4. Sampling of alternatives: model formulation and methodology

As noted previously, the location selection process can be broken into two 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.

The choice set formation process utilizes the concept of an acceptable average commute distance threshold for each household which is based on household and land-use characteristics. This threshold is estimated for each household and used to limit the choices in the choice set. This is done in place of simpler choice set formation methods using heuristics, importance sampling, random selection, etc., to better represent the factors that households likely use when filtering potential household location. An extensive curve fitting exercise was undertaken to find the best distribution representing the critical average work distance at which the households reside, where the average work distance of all employed members of a household to their (potential) residential location is used (see Rashidi and Mohammadian, 2011). It was found that the average work distance follows a Weibull distribution based on the results of the distribution test on work distance using Kolmogorov–Smirnov statistics (Chakravarti et al., 1967; Eadie et al., 1971).

It is assumed that depending on the individual household's attributes, decision makers have some threshold for the maximum commute distance beyond which housing alternatives will not be attractive to the household. In such cases the household will reject any alternative with the distance surpassing the threshold defined for the household. This interpretation of the two continuous dependent variables suggests the use of a hazard-based formulation. In general a hazard model can be formulated as:

$$\lambda(t)dt = \Pr(t + \Delta t \geqslant T \geqslant t|T \geqslant t) = \frac{f(t)dt}{S(t)} = \frac{S'(t)dt}{S(t)} \tag{1}$$

where λ_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 Eq. (1) as:

$$S_{(}t) = \exp\left[-\int_0^t \lambda(u)du\right] \tag{2}$$

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:

$$\lambda_i(d) = \gamma d^{\gamma-1}\exp(-\theta_x X_i) \tag{3}$$

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

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

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

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:

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

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 Eq. (5) are then utilized to generate individual choice sets.

4.1. 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. The frequency of the transit service during the day, especially mid-day, was found significant in the work distance model. This is the only built-environment variable which was found to be statistically significant in the model, while many other land-use variables were tested and not found statistically significant.

In addition, macroeconomic factors like inflation rate and unemployment rate were included in the explanatory variable pool. In order to have all prices and income values be comparable, the first wave of the PSTP that was used 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 Bureau of Labor Statistics (2009).

The average values and standard deviations of the explanatory variables that were found statistically significant in the model are presented in Table 1.

4.2. Modeling results and analysis

The results of the parameter estimation for the choice set formation model are presented in Table 2. Model parameters are estimated by maximizing the likelihood function presented in Eq. (5) using the *nlp* procedure provided by the SAS 9.1.3 statistical package. Before evaluating the quality of the estimated parameters, it should be noted that the effect of covariates in a hazard model on the hazard rate are opposite to the sign on the parameters. In other words, if a covariate has a negative value, the hazard 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.

The Weibull hazard-distribution of the work distance model has a monotonically increasing shape because sigma parameter is greater than one as expected.

It was found that household's current average work distance is considerably affecting the household decision about its new residence. The annual income, which is also positively correlated with the number of vehicles, is also felt to be important in the household choice of the average work distance. Although the parameter is not statistically significant, it shows the expected effect direction and is kept in the model to keep the model sensitive to changes in income for conceptual reasons. The higher the household income is the farther the household members are willing to travel for work. This is largely because

Table 1
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 m by 750 m gridcells.

Table 2
Results of joint model of household average work distance.

Parameter	Estimate	*t* Value	Pr > |*t*|
Household average work distance			
Sigma	1.828	12.505	0.000
Constant	2.847	6.680	0.000
Previous work distance	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 employed	0.193	1.834	0.069
Change in total 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.

wealthier households are also more likely to live in suburban areas and commute further distances. Similarly, total number of vehicles in the household is positively correlated with the work distance. Households with more workers generally commute to farther work destinations, likely due to the need to balance commuting between all of the working household members. Interestingly, households with increases in the number of adults are likely to work closer to their home on average, perhaps reflecting the low value service or retail jobs likely to be held by household members reaching adulthood or moving back into the household (two of the primary reasons for an increase in adults) and conversely an increase in the average distance as these individuals move away. Households living in areas with more available mid-day transit are also more likely to reduce their work distance. Finally, the unemployment rate as a representative of the supply side of the market, was found to be significant in the household average work distance. The results shown in Table 2 imply that any increase in the unemployment rate is related to a households' tendency to reduce the average work distance.

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

The parameter estimates of the model that are presented in Table 2 can then be 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 Eqs. (2)–(4) as:

The probability density function for work distance is:

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

As shown in Eq. (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 Eq. (6). This equation can then generate a probability density function for each household.

5. 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 from which they select their residential locations, because household surveyed in PSTP are in these 741 zones.

Using the probability density function of Eq. (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 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 households, the probability of moving to any of the 741 zones for all the households in the data is defined. This probability is 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, the exponential of the 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 the probability of selecting a zone increases as it gets closer to the job locations of household members while it decreases when households considers 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. The number of alternatives selected for each household is not made fixed, but is randomly realized based on the selection probabilities. Alternatives are selected without replacement for each household and the alternatives with higher probabilities have a greater chance to be selected.

In order to approximate the most appropriate choice set size, nine total random draw values are examined. Table 3 shows the effectiveness of these random draw value scenarios, in terms of how well the scenarios capture the actually selected zones in the choice sets generated.

The first column in Table 3 shows total number of random draws performed for forming the choice sets. 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 in terms of the inclusion of the actual choice in the choice set. 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 predictive ability 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 the actual decision is included in the choice set at an acceptable rate. In other words, increasing choice set size raises the chance of not excluding the decision's maker final choice but increases the complexity of the problem. Behaviorally, it is thought that people generally do not compare a large set of alternatives, instead, a small set of most desirable choices are selected among which the final choice is made. 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. Another important criterion for evaluating a choice set formation method is that the choice set formation method results in consistent parameter estimation. It was found that the choice set formation algorithm of this study results in consistent parameter estimation when it is coupled with the sampling correction method which will be revisited in more detail in the next section.

Fig. 1 shows the tradeoff between the accuracy and complexity of the choice set formation algorithm across different choice set sizes. It can be discerned from Fig. 1 that the algorithm 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 included in the choice set. As a rule of thumb Nerella and Bhat (2004) found that at least one eighth of the universal choice set should be included in the sampled choice set to have consistent parameter estimates during a number of simulation exercises regarding random sampling of alternatives. The modeling results of this study are presented for the case that, on averagely, almost eighteen percent of the universal choice set are included in the choice sets of the individuals, although the samples are not constructed randomly.

Table 3
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

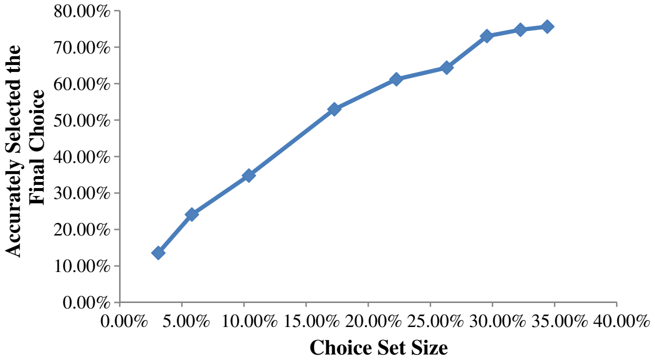

Fig. 1. Tradeoff between choice set formation algorithm accuracy and choice set size.

The inclusion of non-random choice set formation in the model formulation does, however, introduce sampling bias into the subsequent discrete choice model of residential location. It is necessary, then, to diminish the effect of the specific choice set realization on the parameter estimation of the discrete choice model by utilizing a latent correction in the successive model which will be elaborated in more details in the next section.

6. Residential location choice model methodology

Choice set composition can have significant impacts on the results of a discrete choice model. Therefore, it is very critical to account for these effects on the parameter estimates, otherwise the estimated parameters are not consistent. 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 universe of alternatives which are selected probabilistically based on their characteristics. It has been proven (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:

$$P_{ij} = \frac{e^{\mu V_{ij} - \ln C_{ij}}}{\sum_{l=1}^{L} e^{\mu V_{il} - \ln C_{il}}} \tag{7}$$

$$C_{ij} = \frac{q_{ij}}{\sum_{k=1}^{K} q_{ik}} \tag{8}$$

where μ is a scale parameter and V_{ij} is the deterministic utility, K is the total number of alternatives (741) and L is the total number of alternatives in the choice subset. The C_{ij} 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*). 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 Eq. (7), we are assured that the multinomial logit model gives consistent parameter estimates by correcting the sampling bias. Meanwhile attractive alternatives with higher probability than unattractive alternatives are included more frequently among the chosen set of alternatives, reducing the simulation variance.

7. Residential location choice model results

In this section the results of the final multinomial logit model are presented. First, however, a descriptive analysis of the explanatory variables used in this study is undertaken to provide context. Table 4 shows the averages and standard

Table 4
Explanatory variables used in the multinomial logit model.

Parameter	Name	Average	St. Dev.
Log of total number of jobs*	Jobs	4.20	1.99
Log of total number of real estate, rental and leasing jobs**	Real	0.38	0.75
Log of total number of finance and insurance jobs**	Fina	0.43	1.00
Log of number of residential housing units	Unit	2.52	1.11
Log of industrial square feet**	Indsqf	4.94	3.19
Log of manufacturing jobs-neighbors**	Manu_N	3.04	1.54
Log of utility jobs-neighbors**	Util_N	0.31	0.73
Log of total number of finance and insurance jobs-neighbor**	Fina_N	2.65	2.04
Log of government square feet-neighbors**	Govsqft_N	10.49	1.43
Log of number of children (<16)/area***	Child	6.14	1.47
Log of number of middle age (<44 and >35)/area***	Midage	6.20	1.32
Log of number of seniors (<75 and >64)/area***	Senior	5.15	1.54
Absolute difference between average zonal income and HHld income (X100,000) ***	DiffInc	0.23	0.19
Poor X (log of absolute difference between average zonal land value and the average land value of the zone in which HHld lives)***	PoorLandVal	1.44	3.81
Middle X log of absolute difference between average zonal land value and the average land value of the zone in which HHld lives***	MiddleLandVal	7.68	5.35
Rich X log of absolute difference between average zonal land value and the average land value of the zone in which HHld lives***	RichLandVal	2.09	4.42
Transit percentage usage X binary variable for decrease in gas price***	TransitDec	0.07	0.09

* 450 m by 450 m gridcells.
** 750 m by 750 m gridcells.
*** TAZ.

deviations for the independent variables that are used in the multinomial logit model residential location choice model. Many explanatory variables were tested in this study, but only the variables found to be statistically significant in the final model are reported in Table 4.

The explanatory variables used are observed at three geographical resolutions. Land use variables relating to the job type totals in a zone are provided by Puget Sound Regional Council are at a resolution of 750 m by 750 m grid-cells except for the total number of jobs which is at a resolution of 450 m by 450 m gridcells. The rest of the variables (except for land value) are borrowed from CTPP data files which are available at the TAZ level. Land values as discussed earlier were obtained from the assessment department data bank for King and Kitsap Counties. 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 land value for each zone is not used directly in the model, however, as it is transformed into the absolute value of the difference in land value for each TAZ from the current residential location (i.e. the location in the previous wave of the panel). Additionally, three binary variables are defined for three income categories, which are interacted with the land value difference variable, to see how different income classes respond to differences in land values. Households with less than 25,000 annual income are called low-income, household with annual income greater than 75,000 are called high-income and others are called middle. Therefore, there parameters are estimated for three land value difference variables depending on household income.

The first three variables in Table 4 represent the employment densities in the area for total employment and two individual employment categories while the next two variables relate to the residential and industrial land use in the zones. The next four explanatory variables are included in the model to account for spatial dependency between contiguous zones. These four variables represent the land use conditions in the zones surrounding the zone under consideration. Population density was also included in the model, as was density of children and seniors in a TAZ, which can imply whether a TAZ is family oriented or not. The logarithm of the absolute difference between household income and average zonal income was another variable included in the pool of explanatory variable utilized in this study, on the assumption that, much like with land values, households look for zones which are more similar to their socio-demographic attributes. The *DiffInc* variable is designed to capture this effect. Finally, the percentage of transit users in a zone is interacted with a variable which indicates a decrease in the gas price (in real terms) between waves, on the thought that transit oriented areas may be less attractive in an environment of declining fuel prices.

Next, the detailed results of the developed multinomial logit model are presented and discussed. Table 5 shows the estimated parameters of the multinomial logit model with 200 random runs for choice set generation (giving an average choice set size of 128).

General model goodness-of-fit seems very promising based on the results presented on the right hand side of Table 5. Land value as a very critical variable in selecting the zone to which a household decides to move, was found to be a statistically significant variable in the model for all income categories. It can be interpreted from the negative sign of LandVal parameter that zones with greater difference from the land value of the current residential zone become less attractive to the household since they become either less affordable or too affordable, i.e. not having the desired amenities, quality, etc. the household is accustomed to. Decision makers are less interested in zones with higher employment, although this is less the case if those jobs are in the finance industry, likely because these jobs are often in attractive, high value downtown areas. This can be rationalized by the fact that zones with higher employment are not necessarily as oriented toward

Table 5
Multinomial logit model development results.

Parameters	Estimation	*t*-Value
PoorLandVal	−0.268	−7.25
Middle LandVal	−0.2923	−19.25
RichLandVal	−0.2968	−10.58
Correction factor	0.3259	11.44
Jobs	−0.1498	−2.88
Real	−0.5902	−3.31
Fina	0.2207	2.02
Unit	0.1348	2.21
Indsqft	0.041	2.02
Manu_N	−0.1059	−2.96
Util_N	0.1617	1.85
Fina_N	0.0825	1.74
Govsqft_N	−0.1345	−2.35
DiffInc	−0.8824	−2.2
Child	0.2621	2.61
Midage	−0.2246	−1.84
Senior	−0.1574	−2.72
TransitDec	−2.5705	−2.45
Log_likelihood at convergence		−2370
Likelihood ratio		592.95

residential needs or have less attractive qualities due to more traffic, higher densities, and so on. It can be seen in Table 5 that the utility of moving to a zone is magnified if the zone is surrounded by utility and financial industry jobs while it is reduced if it is bordered with zones with manufacturing employment or governmental land-usage. The findings of this study also confirm the intuitive result that zones with greater differences in income relative to the households' income are less attractive. Much like with the land-use results, households tend to move to areas with characteristics similar to their own. The *Seniors* and *Midage* parameters found to be negative which implies that the utility function shrinks if the number of seniors or middle age individuals increases in a zone while the percentage of children has a positive influence. Finally, zones with high transit ridership were found to be less attractive as gas prices decrease. As fuel costs become less of an issue to families they appear to stop focusing as much on transit-oriented zones, which seems reasonable, although the converse of this situation, high auto-dependency interacted with a fuel-price-increase indicator, was not found to be significant.

Finally, it should be noted that although a sample size of average 128 (200 runs) was selected for the final analysis, the modeling results of 43, 77 and 165 (50, 100 and 300 runs) average sample size also showed no more than 42% difference on average between the presented results in Table 5 and the estimated parameters for these three models. In other words, even if other sample sizes had been considered for parameter estimation, the maximum difference between the estimated parameters and the parameters shown in Table 5 would have not exceeded 42%. This clearly shows the importance of including the correction factor in the model which can stabilize the parameter estimates. However if a complete random sample is drawn for each household (100 runs and no correction factor is included) the parameter estimations are at least 300% of what is presented in Table 5. Therefore, it can be concluded that the correction factor can give consistent parameter estimates while also providing a way of including behavioral choice set formation in the discrete choice model.

After finalizing the two stage residential location choice model, which includes both choice set formation and location steps, the results were compared against a simple one-step multinomial location choice model. This analysis was undertaken to estimate the benefits gained from using the more complex choice set formation procedure as previously described. To accomplish this, a second location choice model was estimated using random choice set formation. The model potentially included all of the same explanatory variables included in each stage of the final two-stage model, including average commute distance, land use measures and household level variables. The simplified MNL model was estimated using choice sets of 128 zones which were randomly assigned and the results were compared to the two-stage model. The first comparison was on the overall prediction potential, or the percentage of choice situations where the correct choice was made. Overall, the results show that for choice sets with an average size of 128 the two stage model had a prediction potential of 5.8% against 1.4% for the random choice set model, both of which are significantly higher than the null expectation of 0.7%. It should be noted here that when assessing the prediction potential that it is only the relative differences between the models and the null model that is important rather than the absolute value, as the percent correctly predicted is highly influenced by choice set size. As an illustration of this point, consider that if the choice set is restricted to two zones, even a model based on random choice will have a prediction potential of 50%. Finally a statistic which compares the final estimated average commute distances to the observed values was estimated. The sum of the absolute difference of these distances was calculated for each model and was 7200 in the two-step model and 10,762 in the random choice set model, meaning the two-stage model more accurately recreates the observed average commute distances. So, overall the two-stage model shows improvement over a standard multinomial location choice model using random selection in the choice set formation, as expected.

8. Conclusions and future directions

This study presented a behavioral model of alternative choice set formation for the residential location choice problem as well as its application in a multinomial 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, a choice set is randomly formed, and then from the generated choice set the alternative with the highest random utility can be selected by using traditional discrete choice models. The sampling bias is adjusted in this study by using the sampling of alternatives method that can be found Ben-Akiva and Lerman (1985). An innovative, behavioral sample design method was introduced in this study which uses the household average word distance as the yardstick for evaluating the alternatives. A hazard-based formulation with a Weibull distribution was employed for modeling the sample selection process. During the location choice model estimation process, a choice set was simulated for each decision maker using the choice set formation model. Finally the simulated choice sets were used in a multinomial logit model to model the disaggregate behavior of decision makers in finding a new residential location area. The Puget Sound Transportation Panel of the Seattle Metropolitan Area was used in this study along with other sources of data such as the built environment, land-use, and economic factors. The models developed in this study were validated in several 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.

Future improvements to the model could include: incorporating heterogeneity in the choice set formation, investigating the importance of other variables on housing search choice set formation besides work distance, including the unobserved spatial autocorrelation between the alternatives in the multinomial logit model, and formulating the model based on a more generalized travel cost measure rather than simply the distance to the workplace if new sources of data become available. This last improvement would also have the benefit of weighting the commuting impedance by importance, with the commuting time for higher-income or primary providers impacting more than secondary and part-time workers in formulating the “average commute distance” as done in the current 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 a large number of alternatives need to 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.

Acknowledgments

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.

References

Arnold, S.J., Oum, T.H., Tigert, D.J., 1983. Determinant attributes in retail patronage: seasonal, temporal, regional, and international comparisons. Journal of Marketing Research 20, 149–157.

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

Ben-Akiva, M., Watanatada, T., 1981. Application of continuous choice logit model. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge Mass.

Bushway, S.D., Johnson, B.D., Slocum, L.A., 2007. Is the magic still there? The relevance of the Heckman two-step correction for selection bias in criminology. Journal of Quantitative Criminology 23, 151–178.

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

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

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

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

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

Habib, M.A., Miller, E.J., 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.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

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

Huff, J.O., 1986. Geographic regularities in residential search behavior. Annals of the Association of American Geographers 76, 208–227.

Kanaroglou, P.S., Ferguson, M.R., 1996. Discrete spatial choice models for aggregate destinations. Journal of Regional Science 36 (2), 271–290.

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

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

Kitamura, R., Kostynuil, L., Ting, K.L., 1979. Aggregation in spatial choice modeling. Transportation Science 13, 325–342.

McFadden, D., 1974. Conditional logit analysis on the temporal stability of disaggregate travel demand models. Transportation Research Part B 16, 263–278.

McFadden, D., 1978. Modeling the choice of residential location. In: Karlquist, A., et al. (Eds.), Spatial Interaction Theory and Residential Location. North Holland, Amsterdam, pp. 75–96.

Murakami, E., Watterson, W.T., 1992. The Puget Sound transportation panel after two waves. Transportation 19 (2), 141–158.

Nerella, S., Bhat, C.R., 2004. Numerical analysis of effect of sampling of alternatives in discrete choice models. Transportation Research Record 1894, 11–19.

Rashidi, T.H., Mohammadian, A., Koppelman, F.S., 2011. Modeling interdependencies between vehicle transaction, residential relocation and job change. Transportation 38 (6), 909–932.

Rashidi, T.H., Mohammadian, A., 2011. In: Behavioral Housing Search Choice Set Formation: A Hazard-Based Screening Model of Property Value and Work Distance, ASCE Conf. Proc. T&DI Congress 2011: Integrated Transportation and Development for a Better Tomorrow Proceedings of the First T&DI Congress 2011, doi: 10.1061/41167(398)88.

Salomon, I., Ben-Akiva, M., 1983. The use of the life-cycle concept in travel demand models. Environment and Planning A 15, 623–638.

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

Spiggle, S., Sewall, M.A., 1987. A choice sets model of retail selection. The Journal of Marketing 51 (2), 97–111.

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

Timmermans, H.J.P., Golledge, R.G., 1990. Applications of behavioral research on spatial problems II: preference and choice. Progress in Human, Geography 14, 311–354.

US Bureau of Labor Statistics, 2009. Local Area Unemployment Statistics. <http://www.bls.gov/lau> (last accessed July 2009).

Waddell, P 1996. Accessibility and Residential Location: The Interaction of Workplace, Residential Mobility, and Location Choices, Lincoln Institute of Land Policy TRED Conference, Cambridge, Massachusets.

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, 247–263.

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

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

Wright, P., Barbour, F., 1977. Phased decision strategies. In: Starr, M., Zeleny, M. (Eds.), Management Science. North Holland, Amsterdam, pp. 91–109.

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