Choice Set Formation of Housing Location: A Comparative Analysis

Mehran Fasihozaman Langerudi
PhD student, University of Illinois at Chicago

Mahmoud Javanmardi
PhD student, University of Illinois at Chicago

Kouros Mohammadian, Ph.D.
Associate Professor, University of Illinois at Chicago

P.S Sriraj, Ph.D.
Research Assistant Professor, University of Illinois at Chicago

Behnam Amini, Ph.D.
Assistant Professor, IKIU, Iran

Word Count : 9890

1 Abstract

The problem of choice set formation for decision makers is an important subject in spatial discrete choice modeling especially when the choice set contains a large number of elemental alternatives mostly in spatial modeling applications. In general, the choice set of an individual could be a randomly sampled choice set; however, this is claimed to be a behaviorally inacceptable practice due to the fallacious assumption of full knowledge of individuals about potential random choices. This brings up the need to come up with methods to logically allocate credible choice alternatives for individuals. While the use of these methods could be dependent upon specific applications, this paper attempts to identify the distinction between model estimation and prediction steps in the context of residential location choice modeling that could be studied under spatial econometrics. From a theoretical point of view, the paper proposes a modified weighted stratified sampling approach which is an improved version of spatial random sampling for logit model estimation. It is believed to be a better replicate of the universal choice set than other sampling methods and it is capable of utilizing the maximum information from the sampling space to come up with consistent estimates. Then, the estimated logit model could be applied within a simulation framework using any alternative sampling approach for prediction.

Keywords: Spatial Choice Modeling, Residential Location Choice, Choice Set Formation, Sampling Approach

2 Introduction and Background

Land use planning has long focused on the concept of how households and firms decide where to establish their activities (Alonso 1964), the decisions of where to reside, establish manufacturing sites or opening up retail centers are all among these activities. More specifically, residential location choice modeling is one of the areas in land use planning that attempts to examine household 's location search behavior incorporating their trade-offs between housing quality, prices or rents, distance to work location and other key competing factors (Brown and Moore 1970; Rossi 1955). Later studies on residential location choice modeling go beyond simple models and challenge to develop disaggregate computational models to predict household level housing location decisions. One of the motivations behind these residential choice models is the concept of integrated land use and transportation planning that enables the feedback of land use changes on transportation demand and planning accordingly (Waddell et al. 2007). Such large-scale integrated frameworks would implement micro-simulation of disaggregate units, households as well as their short-term and long-term activities like housing relocation decisions, to examine the impacts of urban development on transportation system. Naming a few of these large-scale microsimulation models, one could refer to ILUTE (Salvini and Miller 2005), UrbanSim (Waddell et al. 2003) and PECAS (Abraham John E. and John D. Hunt 2003).

As mentioned earlier, residential location choice modeling is in the stage of disaggregate computational and econometric models. Researchers have used many different spatial econometric discrete choice models to address this problem. Some studies have focused on single aspects of households ' concerns in residential location choice; for instance, commuting factors in residence choice (Clark and Withers 1999), accessibility to non-work activities (Ben-Akiva and Bowman 1998), travel mode choice (Pinjari et al. 2008a, b) or modeling challenges such as choice set formation in multinomial logit models (MNL) or non-MNLs (Guevara and Ben-Akiva 2013). In terms of different types of discrete choice models, researchers have applied various modeling structures. Kim et al. (2005) developed a nested logit model to deal with large categories of locations. Habib and Kockelman (2008) applied Nested Logit and joint MNL to consider all possible alternatives. Guevara and Ben-Akiva (2006) considered the endogeneity of the unobserved factor with measured variables and tried to address the endogeneity issues through the Berry, Levinsohn and Pakes (BLP) and two-stage least square methods.

A critical step in residential location choice problem through discrete choice modeling is the choice set formation which is associated with the level of geographic aggregation for the alternatives. Even though it is ideal to set parcels of land or buildings as the alternatives that households encounter in choosing housing location (Lee et al. 2010), computational and data availability issues compel researchers towards using more aggregate geographic levels such as neighborhoods or traffic analysis zones (Guo and Bhat 2007). However, the nature of this problem imposes a computational barrier which is the large number of alternatives even in the aggregate case. In order to avoid the infeasible or hardly achievable computational issue of large number of alternatives, researchers have tried different sampling methods to shrink the choice set. It is noteworthy to mention that there are also arguments other than the computational issue against the idea of universal choice set that question the knowledge of individuals about the entire choice possibilities (Fotheringham, 1988).

The primary work in alternative sampling goes back to the precious study of McFadden (1987) who demonstrated that consistent estimation of multinomial logit model can be achieved through random sam-

pling of the alternatives. Many researchers have utilized this result to estimate models in the context of residential location choice (Habib and Miller, 2008). Others have tried to use more behaviorally realistic choice sets to estimate their models, but to account for the bias of the proposed choice sets through sampling bias correction factor (Ben-Akiva and Lerman 1985). This type of sampling is known for importance sampling which assigns more plausible alternatives to the choice set of a household searching for housing (Ben-Akiva and Watanatada, 1981). Following the concept of importance sampling, Zheng and Guo (2008) used a distance constraint for destination location choice to limit the destination choice set of individuals. Auld and Mohammadian (2011) applied time space prism constraint to form the choice set for destination choice in their destination choice model of ADAPTS activity-based model. In another study, Rashidi and Mohammadian (2012) proposed a hazard-based importance sampling approach to reduce the size of choice set by filtering alternatives based on distance to work. In order to assess the performance of sampling methods, Zolfaghari et al (2010) presented a comparison among several common sampling methods like importance sampling with bias correction, importance sampling without bias correction and random sampling.

The following section of the paper describes an introduction to weighted stratified sampling which is demonstrated by simple and practical examples of an effective sampling method which is consistent with universal choice set estimation. The paper then questions the efficiency of sampling bias correction in importance sampling methods and the way they have been employed in the literature and tries to apply the stratified method in MNL model estimation by using importance sampling for prediction purposes.

3 Choice Set Formation for Model Estimation

In estimation of a multinomial logit model, the first step is to define the choice set that an individual would consider in the context of the specific application. In certain applications the choice set size is limited within a number of computationally feasible alternatives; for example, when it comes to estimating a multinomial logit model for travel mode choice to work location, a decision maker would have a few alternatives such as auto, bus, train, bike or multi-modal, etc which should be collectively exhaustive and mutually exclusive. At first glance, it might seem that the exhaustive choice set for all individuals are the listed alternatives; but the concept is very controversial. Imagine the case of an elderly individual who does not consider biking within his/her choice set, therefore presenting an exhaustive choice set for him/her would require dropping an alternative. Should the analyst still include biking as an alternative? One might argue that biking should be included in the model estimation choice set to represent the impact of age on mode choice. The controversy becomes more intense in problems that include large number of alternatives such as housing location and business establishment location. The decision makers in these applications could face hundreds or thousands of alternatives. Although considering the universal choice set in these problems suggests the true likelihood function and therefore the true estimation, researchers try to implement different methodologies to reduce the number of alternatives in the estimation of the logit model maintaining the consistency of parameter estimates.

In this paper we first propose an alternative sampling approach for choice set generation in estimation of a logit model that preserves long range of attribute values as well as consistency in parameter estimates. The approach examines whether other importance sampling methods are unbiased in estimating consistent parameters even after applying sampling bias correction factor. Before presenting the approach, a simplis-

tic scenario in housing location choice behavior is explained to help interpreting the concept behind the approach. Consider a single observation that shows a household has chosen a housing location among several alternatives while the only housing variable at hand to compare the alternatives is the distance to the head of the household's work location. In this example we can assume that the universal choice set consists of 8 alternatives while in a specific sampling approach we reduce the sample size to 4 plausible alternatives that represent more realistic choices. Typically, it is acceptable that households cut their distance to their work locations. As a result, if an importance sampling approach is used to reduce the choice set, the sampled alternatives would have higher probabilities of being chosen; as a result, the closer alternatives will probably stay in the reduced choice set. In this case, lets assume that the distance to work for the eight alternatives(universal choice set) are the ones shown in Table 1; therefore, in order to estimate a logit model based on universal choice set, the likelihood function is maximized to achieve the maximum likelihood estimator according to the following formulas:

$y_{ni} = 1$ if individual n chooses alternative i and 0 otherwise

$L_x(\beta) = \prod_{n=1}^{N} \prod_{i} P_{ni}^{y_{ni}} = \frac{e^{1*\beta}}{e^{0*\beta}+e^{0.5\beta}+e^{\beta}+e^{2\beta}+e^{3\beta}+e^{20\beta}+e^{30\beta}+e^{40\beta}}$

$\partial L_x(\beta)/\partial\beta = 0 \rightarrow$ MAX with R software $\quad \beta$ = -0.296

The negative sign for β shows if the distance to work increases, the probability that the household would choose the housing location decreases. The coefficient is obtained through the real universal choice set and the true likelihood function. Now, suppose an importance sampling strategy is applied to reduce the choice set to 4 alternatives (clearly, this is just an example and 8 alternatives are still few enough to handle computational operations). As mentioned earlier, an importance sampling approach tries to insert more plausible and probable alternatives in to the choice set. In an extreme case here, we assume all the plausible alternatives that are closer to work location are placed in the choice set of the household (* alternatives in Table 1). Then, the likelihood function is set and the maximization is processed.

$L_x(\beta) = \prod_{n=1}^{N} \prod_{i} P_{ni}^{y_{ni}} = \frac{e^{1*\beta}}{e^{0*\beta}+e^{0.5\beta}+e^{\beta}+e^{2\beta}}$ MAX with R software $\quad \beta$ = 0.222

The result is clearly different and the sign of the coefficient this time is positive. It seems that the elimination of the improbable alternatives made the model less sensitive to distance to work variable. Although

Table 1 : Universal Alternatives

Alternative	Distance to Work (miles)	Choice(1 or 0)
1*	0.1	0
2*	0.5	0
3*	1	1
4*	2	0
5	3	0
6	20	0
7	30	0
8	40	0

*Alternatives in the importance sampling approach choice set

the far away housing locations might have the least probability and the household under no circumstance would ever settle in those locations, the existence of them in the sampling choice set of decision makers implies the true impact of variables through the model estimation that turns out in to true parameter estimates. In other words, the expression "Choice Set" is misleading when it comes to estimation of model parameters, even though, a wealthy household might not consider a low-quality neighborhood for living, the fact that it is not considered as an alternative would provide extra information that in turn helps with the correct estimation of model coefficients. This concept is the cornerstone of the proposed methodology that specifically distinguishes between choice set generation for model estimation and alternative generation for prediction or simulation. In other words, model estimation should be performed on a representative sample of the universal choice set while prediction could be performed on a behaviorally acceptable decision-maker specific choice set (a choice set formed by a given decision-maker). Therefore, the realism in choice set formation must be implemented at prediction (simulation) stage rather than model estimation process. For example, in order to predict where a high-income household would reside, the choice set consists of a reasonable higher range of housing prices (i.e., importance sampling) even though the parameters of the model have already been estimated based on the whole range of housing prices (i.e., universal choice set).

However, there are two major arguments in the literature to fix the bias in importance sampling for model estimation. First, researchers try to exclude the sampling variable from the model (Rashidi and Mohammadian 2011); for instance, instead of using distance to work in the model for this case, they try to include other housing variables like housing price, transit accessibility, etc. But, this still carries significant amount of bias because those variables are likely to be correlated with the sampling variable. For example, zones within proximity to a household's work location might represent a neighborhood with higher housing prices; therefore, the choice set is not a representative of the universal choice set. Second, researchers have tried to account for the presence of bias by introducing a sampling bias correction factor. In this case, first a set of plausible alternatives is selected; then, a weight factor is applied to the alternatives in the choice set to preserve the distribution of universal alternatives (Nerella and Bhat 2007). However, one can argue that the non-existence of improbable alternatives may distract the estimates, casting doubt on the consistency of the parameters. The above-mentioned simple empirical example supports random sampling versus any type of importance sampling, which is true as long as the generated random sample covers a reasonable distribution of the universal choice set with respect to the variable under consideration. In other words, the distribution of the random sample with respect to the variable should have the same shape as the distribution of the universal choice set. This would turn out true if the size of the random choice set sample is large enough.

A very fine point is underlined in this problem that requires more explanation to turn our direction from random sampling to a weighted multi-dimensional stratified sampling to maintain the universal choice set distribution of the variable. To get to the point, lets present another simplified example. Assume the setup of the previous example, but the universal choice set for a housing location is as shown in Table 2 where the frequencies of the alternatives are given. If the city has 150 zones and we stratify the zones in to bins with averages of distance to the household head 's work location, as shown in Table 2, there are 10 zones with distance to work of approximately 0.5 mile. It is also shown that the household has chosen one of the 20 zones within roughly one mile distance to the work location. In this case if we write the true likelihood function and try to estimate the true parameter based on the universal choice set we would have:

$$L_x(\beta) = \prod_{n=1}^{N} \prod_{i} P_{ni}^{y_{ni}} = \frac{e^{1*\beta}}{10e^{0.5\beta}+20e^{\beta}+20e^{2\beta}+30e^{3\beta}+40e^{20\beta}+20e^{30\beta}+10e^{40\beta}}$$

MAX with R software β = -1.317

Now, if the reduced choice set sample is obtained through a weighted stratified sampling, meaning that the choice set sample has number of alternatives from each stratum proportional to its real frequency, then the parameter estimate comes out exactly the same as the true estimate. Table 3 shows the weighted stratified sample for the household. Notice that the frequency of the alternatives in the choice set sample is exactly proportional to the real frequency. In this case, the sample consists of 15 alternatives, nonetheless to mention that the choice set includes the real choice of the household. Trying to find the coefficient estimate in this problem through maximum likelihood, we would clearly notice that the likelihood is proportional to the true likelihood function, resulting in the same estimate as the true value. Interestingly, this is the time when we encounter the important topic of Likelihood Principle in Statistics. Based on Likelihood Principle (LP), two likelihood functions are equivalent meaning that they infer the same amount of information about the value of the unknown parameter when they are proportional to each other with a scalar (Berger, J.O. and Wolpert, R.L. 1988).

$$L_x(\beta) = \prod_{n=1}^{N} \prod_{i} P_{ni}^{y_{ni}} = \frac{e^{1*\beta}}{1e^{0.5\beta}+2e^{\beta}+2e^{2\beta}+3e^{3\beta}+4e^{20\beta}+2e^{30\beta}+1e^{40\beta}} = 10 * L_{xuniversal}(\beta) \propto L_{xuniversal}(\beta) \rightarrow \beta = -1.317$$

Table 2 : Universal Alternatives for 150 Alternatives Scenario

Alternative Group	Universal Frequency	Distance to Work (miles)	Choice within Group(1 or 0)
1	10	0.5	0
2	20	1	1
3	20	2	0
4	30	3	0
5	40	20	0
6	20	30	0
7	10	40	0

Table 3 : Weighted Stratified Alternative Sampling

Alternative Group	Sample Frequency	Distance to Work (miles)	Choice within Group(1 or 0)
1	1	0.5	0
2	2	1	1
3	2	2	0
4	3	3	0
5	4	20	0
6	2	30	0
7	1	40	0

It is noteworthy to mention that random sampling is also consistent with the true parameter estimates. However, in case of random sampling we might lose samples from several of the stratums due to the stochastic process that leads to distraction from the true estimate. Therefore, the most efficient approach is proportional sampling out of all of the predefined stratums using the stratum frequency. In this case, the stratum with the least number of alternatives would have at least one member in the choice set sample. If

the universal alternatives are represented with an n-dimentional matrix $A_{i_1 i_2 .. i_k ... i_n}$ in a way that n shows the number of variables used in stratification and i_k is the number of clusters (quantiles) chosen for variable k based on its marginal distribution, then the element $a_{j_1 ... j_n}$ so that $0 < j_k \leq i_k, 0 < k \leq n$ is a collection of all alternatives that have the joint attribute within the corresponding cluster. $|a_{j_1 ... j_n}|$ is the number of elements in this collection which could be in the range of 0 to N(number of universal alternatives) and $\sum_{m_1=1}^{i_1} \sum_{m_1=1}^{i_2} ... \sum_{m_1=1}^{i_n} |a_{m_1 m_2 ... m_n}| = \mathrm{N}$.

Consequently, the minimum number of alternatives to assert that all the collections have at least one alternative in the choice set would be:

$$|a_{min}| = min\{|a_{m_1 .. m_k ... m_n}|, 0 < m_k \leq i_k, 0 < k \leq n\}$$

D_s :Choice set for decision maker s

$$min|D_s| = \sum_{m_1=1}^{i_1} \sum_{m_1=1}^{i_2} ... \sum_{m_1=1}^{i_n} \lceil |a_{m_1 m_2 ... m_n}| / |a_{min}| \rceil \quad (1)$$

In this paper, a weighted multi variable stratified sampling is implemented to form the choice set of individuals. The ideal way in this approach is to stratify the universal alternatives into joint bins of variables that are used in model estimation then extracting samples based on bin frequency. However, in this study due to computational issues, two key independent variables, price of housing and distance to work were selected for joint stratification. For each household the alternatives are categorized into 9 bins representing a table of 3X3 and each variable is divided in to 3-clusters (based on its marginal distribution. Clearly, the joint bins might not have the same number of alternatives and that is why the weight factor comes in to effect. Everything that is mentioned so far, concerns the model estimation process. On the other hand, once the coefficients are obtained, one can use them in a simulation setting for prediction with realistic household-specific alternative sampling. The assumption is that the estimated model through a weighted stratified sampling approach is not biased with household-specific tastes; therefore, it is now possible to use a household-specific alternative sampling in the prediction step. Later in the paper, a hazard-based approach is suggested as an alternative sampling method for prediction.

4 Data

The geographic scope selected for this study is Chicago s 7 county metropolitan area in Northeastern Illinois including Cook, Du Page, Kane, Lake, Kendall, McHenry and Will counties which have a total household population of approximately 2,9 million households covering 1711 Traffic Analysis Zones (TAZ). This study has used the Travel Tracker Survey conducted by Chicago Metropolitan Agency for Planning (CMAP). The survey was designed for the purpose of regional travel demand modeling and included over 10,000 households, providing a detailed travel inventory for the members of each household as well as sociodemographic information. Furthermore, the exact coordinates of home and work location of the households were available to examine various aspects of transportation and land use accessibilities in GIS. The final sample for the study was truncated to about 6000 samples that contained the necessary

information required for the analysis in this work noting that the primary distribution of the sample was preserved in this process. One of the barriers in this study was the static source of data, the fact that the previous housing locations of the households were not available to conceive the pattern behind their movement and as a result the self-selection bias could be a potential issue. However, that should not be a point of concern in this study as the focus of the paper is on choice set formation.

The paper has focused on zonal level residential choice; therefore, TAZs were selected as the zonal level of geography and detailed TAZ level built environment attributes were attained from a number of different data sources. Land Use Inventory of Chicago 2005 was used to extract accessibility measures to land use categories such as Urban Mix, Shopping Malls, and Office. Urban Mix land use category includes retail trade, but neither in shopping malls, office campuses, single structure offices nor hotels according to the Land Use Inventory definition. It basically includes retail trade services such as general merchandise, food, vehicular, eating and drinking places, etc.

CMAP also provided aggregate data for property values in TAZs. A procedure was implemented to obtain average property values for a housing unit based on total housing units in TAZs. Assessment factors were extracted from county assessors website and applied to adjust the property values.

The other source of data provided by CMAP was the number of jobs in various employment categories which could represent the regional job opportunities. Moreover, Census data was used to access zonal demographic information like racial composition. Even though direct TAZ attributes could not be made through Census data, Census tract attributes were assigned to TAZs through spatial join in GIS. By utilizing GIS tools, transportation accessibility and distance to the nearest available transit opportunities were calculated with the help of various transit shapefiles of urban and suburban rail and bus transit system in Chicago region.

Finally, school quality data was extracted from Public School Ranking website (2001) and the index is based on SAT exam score. The crime data factor was accessed through Police Department website, the comprehensive publicly available data (2010) for all the crimes occurred as well as their exact location. A number of main crime types such as homicide, assault, sex offense were selected and simply added to represent the crime level of the TAZs and it was proportioned to the highest number of crimes observed among the TAZs to come up with an index between 0 and 1. Most of the TAZs were ranked in the range under 0.1.

Table 4 demonstrates the summary statistics of the variables used in this study. The first section shows the household level variables and the second part represents the zonal attributes. The mean of the crime index is 0.04, i.e. a considerable number of the TAZs are in the safe zone with few harsh crime occurrences with respect to the worst TAZ in terms of crime frequency. The distance to transit stations are shown in natural logarithm scale and the main variables from which they are extracted were in ft scale. Correlation analysis of the data showed a considerable association between number of cars and distance to rail stations as well as distance to urbanmix and malls. Using the clustering GIS tools, racial clustering among Asian and Black families was also another factor that must be considered in residential location choice problems.

Table 4 : Descriptive Analysis of the Explanatory Variables			
Variable Name	Definition	Mean	SD
Household Level			
NPerson	Number of Persons	2.27	1.27
HHIncome	Household Income ($)	66104	30902
NWorker	Number of Workers	1.25	0.89
NStudent	Number of Students	0.52	0.93
Ncar	Number of Cars	1.64	0.98
LStay	Length of Stay in Current Location (years)	4.1	1.12
Ndriver	Number of Drivers	1.66	0.81
NChild	Number of Children	0.47	0.91
AgeHead	Age of Household Head	54.6	16.02
Zonal Level			
CrimeIn	Crime Level Index between 0(low) and 1 (high)	0.04	0.13
SchoolIn	School Quality Index between 0 and 100	30.29	19.41
LogDisMetra	Log Distance to the nearest Metra station (SuburbanRail)	9.231	0.892
LogDisCTARail	Log Distance to the nearest CTA rail(intra-urban rail) station	8.216	1.146
CTABusstops	Number of CTA Bus(Intra-urban Bus) Stops per sq miles	62.7	49.1
LogDisPACE	Log Distance to the nearest PACE(Suburban Bus) stop	7.867	1.760
ZWhite	Percentage of White people in a TAZ	0.70	0.23
ZBlack	Percentage of Black people in a TAZ	0.14	0.23
ZAsain	Percentage of Asian people in a TAZ	0.05	0.06
AvgValue	Average Housing (unit) Market Value in a TAZ	333510	713396
AvgEmp	Average Total Employment in a TAZ	2433.3	4627.7
DistUrbanMix	Average Distance to UrbanMix Land (ft)	2361	3056
DistMall	Average Distance to Malls	10381	14091
DistOffice	Average Distance to Office Land	9498	11818
lnTAZJOB30	Log number of jobs accessible within 30 minutes drive (am peak)	12.402	1.013

5 Residential Location Choice Model (MNL) Estimation

5.1 Theory

The residential location choice model presented here is estimated with Multinomial Logit Model (MNL). MNL is the most practical approach in various choice-modeling applications. Even though the Irreverence from Independent Alternatives property (IIA) seems to be questionable in the context of residential location choice modeling for MNL, it is still widely used by practitioners due to the fact that MNL has closed formulation for probability calculation that makes it computationally efficient and readily interpretable as well as the possibility of consistent estimation through a sampling of alternatives. However, as explained in the previous section, consistent estimation can be efficiently obtained through a weighted stratified sampling rather than any importance sampling method coupled with bias correction. It was stated that importance sampling would remove the bias from the sampled alternatives through bias correction but the non-existence of the unobserved samples is the obscurity in consistent estimation. The typical formula for

a MNL model with bias correction factor is (Ben-Akiva and Lerman 1985):

D_n :Choice set for decision maker n
P_{nj} : probability that decision maker n chooses alternative j

$$P_{nj} = \frac{e^{V_{nj}+ln\pi(D_n|j)}}{\sum_{i\in D_n} e^{V_{ni}+ln\pi(D_n|i)}}$$

The bias correction term refers to the probability that the decision maker would consider a specific choice set given the real alternative selected. In case of importance sampling, the calculation of the bias correction seems controversial; therefore, a weighted stratified sampling was applied to replicate the pattern of the true likelihood function.

5.2 Weighted Stratified Sampling

As mentioned earlier the weighted stratification should be implemented on all the variables used in the model to divide the alternatives to categories with joint attributes. However, in order to make the computational work of this problem fair enough, two key attributes, average property value in TAZs and distance to work were selected. The hypothesis was that the two attributes are independent and they can finely stratify the alternatives for the households. Universal alternatives were then divided in to 9 categories, a joint categorization of the two attributes, each one with 3 clusters. The clusters for average property value in the TAZs are not dependent on households. Unlike property value, distance to work clusters are dependent on the household head member 's workplace; thus, the clusters were calculated for each TAZ work place. According to the formula (1), the number of alternatives from each category in the choice set of a household was selected proportional to the number of TAZs in each category. In this study, with the assumption of 2 alternatives being extracted from the category with minimum size, the number of alternatives from other categories could be proportionally calculated. Then, the samples from each category were randomly selected. The average choice set size that was considered for the households with this assumption was 28.

5.3 Estimation

The residential location choice model was estimated with MNL approach and the weighted stratified sampling method. The utility function for zonal level residential choice models is generally a combination of zonal attributes along with interactive zonal and socioeconomic household variables which would take the form of:

$U_{ni} = \sum \beta_j Z_{ji} + \sum \beta_m f(Z_{mi}, H_{mn})$
Z_{ji} is the j-th attribute of alternative i
H_{mn} is the m-th attribute of household n

The interaction function could take various formulations; for example, logarithm of a zonal attribute like distance to the closest rail station multiplied by a binary household variable that indicates whether the household has a car or not. The role of interactive attributes is significant since these attributes represent household tastes and heterogeneity in housing search behavior. On the other hand, zonal variables lack the capability of explaining household-specific taste variations. It is also important to remind that household variables could not be used without interaction with zonal variables since they do not vary across zonal alternatives.

Selection of the variables used in this model was made based on the likely determinants of household behavior in housing location choice discussed in the literature as well as their significance level in model estimation results. The following socio demographic and zonal attributes have been used in the final estimated model:

Household Level:

- Income
- Race
- Number of cars
- Number of children
- Rent or own the house (binary)
- Education
- Number of workers
- Age of the household head member

Zonal Level:

- Zonal crime index
- Zonal school quality index
- Percentage of White households in the Zone
- Percentage of Black households in the Zone
- Percentage of Asian households in the Zone
- Average property value in the zone
- Distance to the closest rail station
- Distance to the local suburban bus transit system (PACE)
- Number of bus stops within the zone per square mile (within city CTA bus)
- Zone is within urban area or suburban area
- Distance to the household 's work location
- Total zonal employment
- Zone center 's distance to the closest UrbanMix landuse type
- Zone center 's distance to the closest RetailMall landuse type
- Zone center 's distance to the closest Office landuse type

Table 5 illustrates the MNL model estimated parameters as well as goodness of fit result for the residential location choice problem for Chicago metropolitan area. As noted before, due to lack of a panel data or dynamic sample, the model would suffer from potential self-selection bias; however, the variety of attributes used to explore the interdependencies between household socioeconomic features and their spatial location choice could compensate for the missing information. A descriptive analysis of the sample revealed extensive information for undertaking appropriate interactive zonal and household variables. These variables were mostly constructed as a production of a binary household statement and a zonal variable, i.e. they represent household-specific preferences with regard to spatial attachments. Looking at Table 5, it is noticeable that school quality is a determining factor for households with higher-average income. Nonetheless, this variable was not significant for rest of the households; hence, it was omitted from the model. As expected, crime level is another distinguishable factor in the model; households, under any circumstances, tend to prefer living in safer neighborhoods. Racial composition of the zones is the next determining motivation for households with different racial backgrounds to select housing location. It is interesting that racial clustering behavior is vivid among Asian and Black households and the fact that they strongly have the tendency to live in neighborhoods with higher percentage of their own races while White households are more flexible, they rather reside in zones where the percentage of Caucasian households is not less than %30 of the total households. They do not necessarily look for zones with higher White percentages; despite the fact that this might be due to the larger population of White households. Unfortunately, the dataset did not record Hispanic as a separate racial background; accordingly, the analysis could not be executed for them. Considering the next variable, property value over income which is appeared as two separate binary variables, suggests that households try to keep this fraction within an affordable but still satisfactory range; as a result, two binary variables were constructed to distinguish whether the average property value of an alternative is too high or too low with respect to the household income. The range cut-offs were obtained from the tales of the variable distribution over the sample.

The impact of transportation accessibility measures were also tested in household location choice. Both intuitively and statistically, household car availability reduces the dependency on public transit allowing the households to expand distances from transit access points. Within the city of Chicago, households with no car tend to live in the zones with higher bus stop density which is apparent in the estimated coefficient. For suburban bus stops (PACE), however, households who own cars, happen to keep their distance from the bus stops. On the other hand, in general, households prefer to reside closer to rail stations. Having said that, households with no car are even more conscious about staying closer to rail access points which is represented in the corresponding coefficients.

Moreover, other household attributes are also playing their role in location choice. Having children encourages households to live in suburban areas. If a household is looking to rent, it is more likely to settle within the city rather than suburbs. Contrary to the common belief that baby boomers nowadays prefer to move to the city and CBD area, at least in the context of this study, the static sample detects association of age with suburban areas. Furthermore, long distances to work, urbanmix and shopping mall land-use types as well as high property values for low-income households are among negative factors in selecting residential location. Finally, the model suggests that higher employment rates as well as higher distances to office land-use type are desirable factors in residential location choice.

The goodness of fit result that is shown in Table 5 show a Rho-square factor is 0.11 which is typically an acceptable value for a spatial location choice problem considering the large number of alternatives. According to the Wilks theorem (Wilks, S. S. 1938), the distribution of the statistic $-2ln(L_0/L_{mle})$ is

asymptotically Chi-square, in this case with 22 degrees of freedom. With the null hypothesis (constant model), the value of the statistic becomes very large $(-2 \times (-20195 + 17947)) = 4496$ and therefore we can reject the null hypothesis in favor of the maximum likelihood estimator (MLE) with very high significance (%99) level. The AIC statistic is also calculated; however, this statistic is usually appropriate for comparing two models. AIC for this model is 35938 compared to 40390 for the null model (constant). Smaller values of AIC represent higher likelihood ratio and equivalently lead to better models.

Table 5 : Residential Location Choice MNL Estimation

Variable Name	Definition	Estimate	t-value
CRIMEIN	Zonal Crime Index	-0.16	-2.22
SCHHIGH	Zonal School Quality × (HH Income >average Income)*	0.018	9.36
WHITLES30	(Zonal White Percentage <%30) × (HH = White)	-1.423	-12.85
ASIANHHZ	Zonal Asian Percentage × (HH = Asian)	8.037	7.9
BLACKHHZ	Zonal Black Percentage × (HH = Black)	2.408	17.31
PRICOINC1	(Average Zonal Property Value / Income <1.6)	-0.205	-3.58
PRICOINC3	(Average Zonal Property Value / Income >5)	-0.790	-15.60
CTACNCAR	CTAbusstops × (Cars =0)	0.002	1.94
PACECAR	LogDisPACE × (Cars >0)	0.103	15.66
SUBCHILD	(Suburban Zone)× (Children >0)	0.16	1.97
CITYAGE	(Zone within the city) × household head's age	-0.007	-4.89
RENTCITY	(Rent)× (Zone within city)	0.542	5.51
DISTAWOR	Zone Center's Distance to HH Work Location	-0.05	-28.06
RAILNCAR	Log Zone Center's distance to the closest rail station × (car = 0)	-0.276	-4.6
RAILCAR	Log Zone Center's distance to the closest rail station × (car >0)	-0.138	-7.76
PRICFORH	Average Property Value × (Income >100k) × D-06	0.889	-5.37
PRICFORL	Average Property Value × (Income <25k) × D-05	-0.342	-4.57
LGPRXMIX	Log Average distance to Urbanmix (ft)	-0.78	-12.39
LGPRXOFF	Log Average distance to Office landuse type (ft)	0.84	7.28
PRXRET2M	(Distance to the closest Shopping Mall >2 miles)	-0.141	4.62
EDUCPRIC	Average Property Value × (HH head member's educaion level is less than college degree) × D-05	-0.360	-6.51
TOTEMPWO	Total number of jobs × number of workers in the household ×D-05	0.424	3.24
	Summary Statistics		
	Number of observations	6047	
	Average choice set size	28	
	Log likelihood function for null hypothesis (constant)	-20195	
	Log likelihood function	-17947	
	R-square	0.11	
	AIC	35938	

* The parts in Parentheses are binary 1 or 0 statements.

6 Hazard Models for Prediction Step

6.1 Theory

As stated earlier, it is still a gap in the literature on the possibility of utilizing a different methodology for choice set formation in the prediction step. It is always desirable to estimate a reliable model that has capability of prediction over different household choice sets. The claim that choice set formation methodology for prediction should be implemented on exactly the same method as estimation, is very controversial and requires more research both from statistical and behavioral point of view.

In order to combine statistical and behavioral soundness, two hazard-based models were developed to justify the behavioral choice set of the households for prediction. The hazard-based models have the potential to constrain household choices based on two criteria, affordability and maximum distance to work thresholds. Based on household socioeconomic attributes, households are assumed to have particular tolerance level for affordability and distance thresholds which are obtained with the help of hazard models. According to the previous studies (Rashidi et al, 2012, Zolfaghari et al. 2010), Weibull distribution was selected for probability distribution of housing price and distance to work for households. Typically, the scale parameter in Weibull distribution specifies the role of socioeconomic characteristics of a household in the probability distribution of price and distance to work. The hazard and density functions for Weibull distribution is according to the following formulas:

$$\lambda(x) = \alpha\beta x^{\beta-1} \quad f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}$$

Where α, the scale parameter, is reparameterized as a function of socioeconomic attributes of the households to determine the heterogeneity of households with respect to these two hazard variables. As a result, the proportional hazard and density functions for housing price and distance to work would look like:

Scale parameter : $\alpha = e^{\theta_0 - \hat{\theta}\hat{X}}$

d: Distance to Work p: Housing Price

$$f(d) = \beta e^{(\theta_0 - \hat{\theta}\hat{X})} d^{\beta-1} e^{-e^{(\theta_0 - \hat{\theta}\hat{X})} d^{\beta}}$$

$$f(p) = \beta e^{(\theta_0 - \hat{\theta}\hat{X})} p^{\beta-1} e^{-e^{(\theta_0 - \hat{\theta}\hat{X})} p^{\beta}}$$

Using the probability density functions, two separate likelihood functions can be written to estimate the household-specific coefficients according to:

$$Ł_d = \prod_{i=1}^{n} f_i(d)$$

$$Ł_p = \prod_{i=1}^{n} f_i(p)$$

A number of reasonable and intuitive variables were selected to estimate both of the models. The deterministic thresholds jointly used in prediction, are explained later in the result section.

6.2 Estimation

Table 6 shows the result of the hazard models developed for housing price and distance to work. For distance to work model, income, car ownership and age of household's head member are the key determinant of the hazard equation. It is noteworthy to remind that the covariates in the model are formulated with a negative sign. To interpret that, for example, income parameter has come out positive meaning that its effect on the hazard function is negative. Therefore, higher income reduces the hazard rate i.e. affluent households are probabilistically inclined to housing with higher property values as well as housing further away from their work location. Car ownership is the next determinant in making longer distances to work location possible. On the other hand, age has negative association with distance which means households with elderly head members tend to locate closer to their work location. For property price, income is the motivation for higher property prices while number of workers and number of children increase the hazard rate towards lower property prices. It could be concluded that between two households with the same income, all being the same, the one with higher number of children and more workers would select a less pricy house.

Log likelihood functions for constant model along with log likelihood measure for the MLE are given for both models. For the distance to work model, likelihood ratio statistic equals 32. In order to assert that the model with household-specific variables works better than a model with just constants (just parameters θ_0 and β), the null hypothesis of the constants must be tested. Based on wilks theorem, the likelihood ratio statistic must have a chi-square distribution with 3 (=5-2) degrees of freedom. The likelihood ratio statistic for this model is 32 which is placed at the very end of the right tale of this chi-square distribution letting us to reject the null in favor of the model with 99% significance level. The same justification could be made in favor of the property price model which has a likelihood ratio statistic of 418.

7 Results

The idea that choice set formation could be different in estimation and prediction is the motivation to put forward behavioral approaches that make the choice set more compatible with intuition as well as to preserve the consistency of the estimates. The first section of this paper claimed that the weighted stratified sampling is not only consistent but also it covers more information for inference of the unknown parameters by checking out a comprehensive range of alternatives. Noticing that estimation procedure should follow a weighted stratified sampling or random pattern in case of relatively large choice sets, researchers still need to take in to account the intuitive choice of alternatives by decision makers through prediction step. In order to try this method, an intuitive hazard-based approach was implemented for

Table 6 : Hazard Models		
Distance to Work Model		
Parameter	Estimate	t-value
θ_0	-0.42	-6.77
β	0.26	81.16
Income ($\times 1/1000$)	0.001	1.96
Number of Cars	0.025	1.76
Household Age of the head	-0.004	-4.15
Summary Statistics		
Number of Observations	6047	
log likelihood function for null hypothesis (just θ_0, β)	-8625	
log likelihood function	-8609	
-2($\ell(0) - \ell(\theta)$)	32	
Housing Price Model		
θ_0	-13.45	-129.8
β	1.11	146.8
Income ($\times 1/1000$)	0.01	20.18
Number of Workers	-0.093	-4.43
Number of Children	-0.11	-7.64
Summary Statistics		
Number of Observations	6047	
log likelihood function for null hypothesis (just θ_0, β)	-82873	
log likelihood function	-82664	
-2($\ell(0) - \ell(\theta)$)	418	

predicting the location choice of a hold-out sample and it was compared to the prediction based on the weighted stratified sampling. It is noteworthy to remind that the models used for both predictions are estimated through the weighted stratified sampling. The performance of the predictions is explored by comparing the results to the actual location choices of the hold-out sample.

The intuitive hazard-based approach is comprised of two independent hazard models for acceptable housing price and distance to work as explained in the previous section. These models were estimated to be applied in limiting the choice set of individuals for more behavioral and realistic choices relative to household socioeconomic conditions. In other words, the models introduce household-specific hazard and probability density functions for acceptable housing price and distance to work. Based on the probability density functions, one can filter the less probable choices through various filtering methods such as cut-offs through cumulative density function (cdf). The filtering used in this method was based on %80 cdf for distance to work and the middle %80 interval of acceptable housing price cdf displayed as the following:

$\alpha_{\%80}$: maximum acceptable distance to work $F(x<\alpha_{\%80}) = \%80$
$\nu_{\%90}$: maximum acceptable housing price $\quad \nu_{\%10}$: minimum acceptable housing price
$F(\nu_{\%10}<p<\nu_{\%90}) = \%80$

The choice of percentage cut-offs are very crucial and extreme cut-offs could eliminate a large portion of alternatives. On the other hand, very conservative cut-offs might not represent the real choice set forma-

tion behavior. Having said that, range of %80 coverage was found as an acceptable balance of the two factors not to mention the significant need for research in this arena. Since the hold-out sample was small including approximately 2000 households, comparing zonal re-location of the households to their actual zone residence was not rational; therefore, to rationalize the process the result of the zonal predictions were aggregated to sub-counties in suburban counties and neighborhoods within city limits. For representing the performance of the predictions, looking for the re-location zone assigned to each household to be exactly the same as their actual zone, is neither feasible nor logical; however, comparing the similarities between them is an acceptable comparison strategy. Consequently, Figure 1 is displayed to show the distribution of median income through both prediction methods and their comparison to the actual distribution geographic profile. Comparison between the weighted stratified and hazard-based predictions are hardly achievable by just looking at the figure. As a result, to facilitate the comparison, root mean squared error (RMSE), and average relative error between the prediction result and the actual profile were calculated to show the scale of the induced errors. Table 7 shows the RMSE and relative error for certain household socioeconomic variables aggregated for sub-county and neighborhood level of geography. RMSE and relative error are calculated based on the formulations below:

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum(\hat{X}_i - X_i)^2}$$

i: Each single subcounty or neighborhood within Chicago seven county area that contain assigend samples in actual and predicted conditions

X_i :An actual household attribute $\hat{X}_i$:A predicted household attribute

Relative Error for non-zero attributes: $\eta = |\frac{\hat{X}_i - X_i}{X_i}|$

Table 7 : RSME and Relative Error

Variable	$RSME_1$	$RSME_2$	η_1	η_2
Number of Persons	0.83	0.75	0.24	0.20
Household Income $	19231	16869	0.17	0.12
Number of workers	0.50	0.64	-	-
Number of students	0.82	0.72	-	-
Number of cars	0.60	0.47	-	-
Number of drivers	0.49	0.55	-	-
Number of children	0.80	0.46	-	-
Household Age of Head	7.26	8.5	0.11	0.14

1.Weighted Stratified Sampling for Estimation + Weighted Stratified Sampling for Prediction

2.Weighted Stratified Sampling for Estimation + Hazard-based Sampling for Prediction

The table shows that for most of the attributes the relative error decreases when hazard models are used for prediction choice set. One of the key attributes is income distribution throughout the geography which is the best determinant for prediction. The result shows that the error in income distribution is %5 less when an intuitive approach is used for prediction. Moreover, the error in most of the other variables including number of persons, students, cars, children, income and distance to work has decreased. Even though one example might not be sufficient to get to a broad conclusion, the result brings about the differences between the effectiveness choice set formation for model prediction vs. estimation. It is the fact that estimation is the statistical process that requires maximum information and a broad choice of alternatives while prediction is the intuitive process that should get along with common sense.

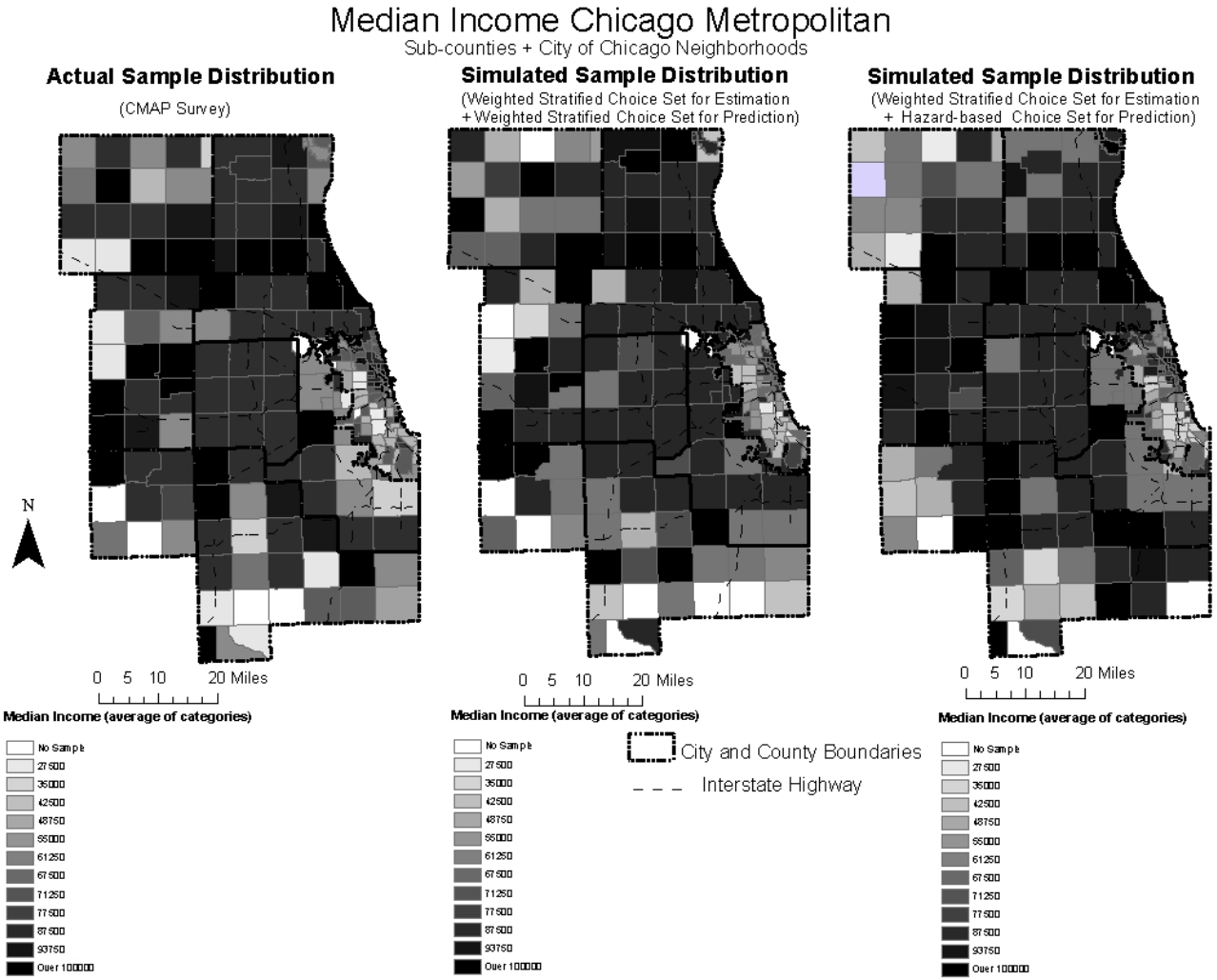

Figure 1: Comparison of median income distribution profile for the prediction approaches

8 Conclusion

This study was conducted to target an important research gap in the context of alternative sampling for choice set formation. Even though the idea of bias correction for sampling is very common in the literature, this study tried to shed light on the potential issues involved in bias correction for importance sampling strategies and attempts to promote the idea that unbiased estimators are hardly achievable through importance sampling even after applying bias correction weights. It was explained that the existence of unobservedalternatives in the estimation choice set would cover a variety of alternatives leading to more informative coefficient estimates. In order to replicate an alternative sampling strategy, with likelihood function proportional to the universal likelihood function, the weighted stratified sampling method was introduced to preserve the real distribution of the covariates used in model estimation. The estimated model through weighted stratified sampling could now be a standpoint for prediction.
On the other hand, the true and rational choice set of households, when it comes to prediction, is conditional to their socioeconomic status. Therefore, hazard-based models were applied to filter conditional alternatives for prediction step of a previously estimated model. The results of prediction for this two-step model were compared with the results generated from a model estimated and predicted with the same sampling method. It was proved that such a two-step model has a better prediction potential as well as a more realistic statistical justification. For future work, it is valuable to improve the prediction potential with more realistic alternative sampling approaches in prediction step to account for more detailed household socioeconomic conditions along with housing supply information to avoid over-concentration of households in certain geographic locations. Interestingly, the hazard-models that are used for the prediction step of this model eliminated a small portion (%20) of improbable alternatives through the tail of distance to work and price of housing probability density functions; however, due to the few number of observations used to predict the model, the issue of over-concentration was not a significant problem. However, for large number of observations, fine consideration of housing supply and capacity constraints must be implemented to balance out the prediction results as the prediction choice set strategy becomes more behavioral.

9 References

Alonso, W.: Location and Land Use: Toward a General Theory of Land Rent. Harvard University Press, Cambridge (1964)

Auld, J., Mohammadian, A.: Planning constrained destination choice in the ADAPTS activity-based model. Paper presented at the 90th annual transportation research board meeting, Washington DC (2011)

Ben-Akiva, M.E., Bowman, J.L.: Integration of an activity-based model system and a residential location model. Urban Stud. 35, 11311153 (1998)

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

Ben-Akiva, M. , Watanatada, T. 1981. Application of a continuous spatial choice logit model. Struc-

tural analysis of discrete data with econometric applications. MIT Press, Cambridge, Mass.

Berger, J.O., Wolpert, R.L. The Likelihood Principle (2nd ed.). Haywood, CA: The Institute of Mathematical Statistics. ISBN 0-940600-13-7.(1988)

Brown, L.A., Moore, E.G.: The intra-urban migration process: a perspective. Geogr. Ann. B 52, 113 (1970)

Clark, W.A.V., Withers, S.D.: Changing jobs and changing houses: mobility outcomes of employment transitions. J. Regional Sci. 39, 653673 (1999)

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

Guevara, C.A., Ben-Akiva, M.E.: Endogeneity in residential location choice models. Transp. Res. Record 1977, 6066 (2006)

Guevara, C.A., Ben-Akiva, M.E.: Sampling of alternatives in Multivariate Extreme Value (MEV) models, Transportation Research Part B: Methodological, 31-52(2013)

Guo, J.Y., Bhat, C.R.: Operationalizing the concept of neighborhood: application to residential location choice analysis. J. Transp. Geogr. 15, 3145 (2007)

Habib, K.M.N., Kockelman, K.: Modeling choice of residential location and home type: recent movers in Austin Texas. In: 87th Annual Meeting of the Transportation Research Board, Washington, DC (2008)

Habib, M.A., Miller, E.J.: Modeling residential mobility and spatial search behaviorestimation of continuous-time hazard and discrete-time panel logit models for residential mobility. In: Transportation Research Board 87th Annual Meeting, Washington, DC (2008)

Hunt, J.D., Abraham, J.E. (2003) Design and application of the PECAS land use modelling system. Paper presented at the 8th International Conference on Computers in Urban Panning and Urban Management, Sendai, Japan.

Lee, B.H.Y., Waddell, P.A., Wang, L., Pendyala, R.M.: Operationalizing time-space prism accessibility in a building-level residential choice model: empirical results from the Puget Sound region. Environ. Plann.A 42 (2010)

Kim, J., Pagliara, F., Preston, J.: An analysis of residential location choice behaviour in Oxfordshire UKa combined state preference approach. Int. Rev. Public Admin. 8, 103114 (2003)

McFadden, D.: Modelling the choice of residential location. In: Karlqvist, A., Lundqvist, L., Snickars, F.,Weibull, J. (eds.) Spatial Interaction Theory and Planning Models, pp. 7596. North Holland, Amsterdam (1978)

Nerella, Sriharsha, and Chandra R. Bhat. "Numerical analysis of effect of sampling of alternatives in

discrete choice models." Transportation Research Record: Journal of the Transportation Research Board 1894.1 (2004): 11-19.

Pinjari, A.R., Eluru, N., Bhat, C.R., Pendyala, R.M., Spissu, E.: Joint model of choice of residential neighborhood and bicycle ownership: accounting for self-selection and unobserved heterogeneity. Transp. Res. Record 2082, 1726 (2008a)

Pinjari, A.R., Pendyala, R.M., Bhat, C.R., Waddell, P.A.: Modeling the choice continuum: Integrated modelof residential location automobile ownership bicycle ownership and commute tour mode choice decisions. In: Transportation Research Board 87th Annual Meeting, Washington, DC (2008b)

Rashidi, T. H., J. Auld, and A. Mohammadian. A Behavioral Housing Search Model: Two-Stage Hazard-Based and Multinomial Logit Approach to Choice Formation and Location Selection, Transportation Research Part A 46:1097107.(2012)

Rossi, P.H.: Why Families Move: A Study in the Social Psychology of Urban Residential Mobility. Free Press, Glencoe (1955)

Salvini, P.A., Miller, E.J.: ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems. Netw. Spatial Econ. 5, 217234 (2005)

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M., Ulfarsson, G.: Microsimulation of urban development and location choices: design and implementation of UrbanSim. Netw. Spatial Econ. 3, 4367 (2003)

Waddell, P., Ulfarsson, G.F., Franklin, J.P., Lobb, J.: Incorporating land use in metropolitan transportation planning. Transp. Res. A 41, 382410 (2007)

Zheng, J. , Guo, J. 2008. Destination Choice Model Incorporating Choice Set Formation,In: Transportation Research Board 87th Annual Meeting, Washington, DC (2008)

Zolfaghari A., Sivakumar A., Polak J.,Choice set formation in residential location choice modelling: implementation of a hazard-based approach,In:International Choice Modelling Conference (2011)