Choice Set Formation in Residential Location Choice Modelling: Empirical Comparison of Alternative Approaches

Alireza Zolfaghari
PhD Student
Imperial College London

Aruna Sivakumar
Post-doctorate Research Fellow
Imperial College London

John W. Polak
Professor of Transport Demand, Head of Centre for Transport Studies
Imperial College London
ABSTRACT

The discrete choice analysis of residential location choice forms an important part of land use-transport modelling systems but gives rise to a number of significant modelling challenges, one of which is the choice set formation problem. A number of alternative approaches exist to addressing the choice set formation problem, but to the best of our knowledge, there have been no efforts to empirically compare the performance of alternative choice set formation techniques within the context of modelling residential location. In this paper, we propose to address this gap by examining the performance of several choice set formation methods within the context of residential location choice in London. In particular, we implement a recently developed hazard-based model of housing search choice set formation (Rashidi and Mohammadian, 2010). The hazard-based screening model is implemented in two different ways: (a) using deterministic thresholds on different attributes, and (b) using a probabilistic, importance sampling approach. Moreover, the structure of the hazard-based screening model enables us to model systematic and random inter-individual heterogeneity in choice set formation.

The paper presents a brief review of the relevant theoretical and empirical literature on the treatment of the choice set formation problem in the context of spatial choice and sets out the characteristics of the specific modelling approaches implemented in the empirical research. The empirical research uses a specially constructed dataset on residential location choice, developed by the authors for the Greater London area, which draws on information from a number of separate data sources (including revealed preference data from the London Area Travel Survey). The alternative choice set formation approaches are assessed in terms of their prediction performance on a hold-out validation sub-sample and also in terms of estimated model parameters.

Keywords: Residential Location Choice Model, Choice Set Formation, Importance Sampling, Housing Screening Model
1. INTRODUCTION

Models of residential relocation and location choice play an important role in urban planning and policy-making. On the one hand, they help us understand how residential location choices are made; how and to what extent such factors as accessibility, distance from work, price, school quality and safety concerns have an impact on where people choose to live (see for ex. Ben-Akiva and Bowman, 1998, Guo and Bhat, 2004). On the other hand, residential location choice models are an important component of integrated land use-transport model systems, as they predict the medium-term dynamics in the urban area and help to determine how the urban landscape is shaped over time (Waddell et al., 2003).

There are many issues to be explored within the context of spatial models of residential location choice, such as spatial correlation, choice set formation, relocation decisions, the role of accessibility etc. In this paper we focus on the choice set formation issue. Conceptually, residential location choice is a dynamic spatial search process, in which households are exposed to a dynamically changing set of residential alternatives from which, at any point in time, they assemble and evaluate a choice set of credible alternatives, and ultimately, at some point in time, make a selection (Habib and Miller, 2007). However, this underlying dynamic spatial search and choice process is typically unobserved. Without direct observations of how choice sets are dynamically assembled, analysts either face the task of dealing with extremely large universal choice sets (typically numbering many hundreds or thousands of elementary locational alternatives) or they must develop methods of pruning these large choice sets.

In broad terms, two types of approach have been proposed in the literature. In the first approach, the issues of dynamic search are ignored and it is assumed that households do indeed choose from the underlying universal choice set but the computational burden of working with such large choice sets is avoided by adopting various statistical strategies for sampling or pruning the global choice set (e.g., choice sets formed by random sampling of the global choice set) (McFadden, 1978). In the second approach, the existence of an underlying dynamic search process is acknowledged and attempts are made to formulate behaviourally plausible rules describing the operation of aspects of this underlying search process and use these rules to reduce the universal choice set to manageable proportions (Fotheringham, 1988). Sometimes elements of these two approaches are combined (e.g., in route choice context importance sampling approach used to generate choice set), as in the work of (Frejinger et al., 2009). However, to date, there appears to have been no attempt to evaluate, using real residential location choice data, the empirical performance of these different approaches. This is an important gap in the literature, which this paper aims to address.

In particular, we compare empirically the uniform random sampling approach with a deterministic choice set formation approach, based on a hazard-based housing screening model developed by Rashidi and Mohammadian (2010) and compare each again with the results obtained using the universal choice set. The hazard-based screening model is implemented using a deterministic approach as well as an importance sampling approach with and without bias correction (more details in section 3.2). The residential choice model is estimated as a multinomial logit (MNL) choice model. Models are compared based on how well they predict on a hold-out validation sample.
The rest of this paper is organised as follows. Section 2 reviews the literature in spatial location choice and choice set formation, focusing primarily on the residential location choice context. Section 3 presents the modelling framework for this research, including the details of the choice set formation models that were implemented. Section 4 presents the details of the empirical analysis, including the model estimation and validation results and a discussion of the findings. Section 5 concludes this paper.

2. LITERATURE REVIEW

There is a large body of literature on modelling residential location choice, ranging from the early works of McFadden (1978) to Guo (2004) who focuses on spatial complexities in residential location choice models, Bayoh et al. (2006) who examine the determinants of residential location choice, Guevara and Ben-Akiva (2006) who study endogeneity issues, Habib and Miller (2009) who analyse reference-dependence within a relocation context, and the recent compilation of Pagliara and Wilson (2010) which presents residential location models that are implemented today. In this section, we focus on the literature devoted to spatial choice set formation in general, and within the context of residential location choice modelling in particular.

In spatial choice situations, decision makers ostensibly face a very large set of alternatives. However, estimation of location choice models with the entire spatial choice set that is in principle available to each household is not desirable. Firstly, it is computationally burdensome despite recent advances in computational power; depending on the granularity of the location choices, the universal choice set can range from hundreds (zone-based location choice) to hundreds of thousands (parcel-based location choice model). Secondly, such a model is not behaviourally realistic, as individuals and households do not actually consider the entire set of spatial alternatives in making their choice (Fotheringham, 1988). The actual evaluated number of alternatives will be constrained by the dynamics of the supply of alternatives (e.g., the structure of the residential location market) and by individuals’ limited capacity for gathering and processing information (Fotheringham et al., 2000). Shocker et al. (1991) suggest that individuals make decisions based upon four hierarchical or nested sets of alternatives: the universal set, awareness set, consideration set and choice set.

<table>
<thead>
<tr>
<th>Choice set formation approach</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform Random Sampling of Alternatives</td>
<td>Most of the studies in residential location choice modeling: Habib and Miller, 2009; Lee and Waddell, 2010</td>
</tr>
<tr>
<td>Importance Sampling of Alternatives not considering heterogeneity of decision makers</td>
<td>Destination choice: Ben-Akiva and Watanatada, 1981 Route choice: Frejinger et al., 2009</td>
</tr>
<tr>
<td>Deterministic constraint not considering heterogeneity of decision makers</td>
<td>Destination choice: Termansen et al., 2004; Scott, 2006</td>
</tr>
<tr>
<td>Stochastic constraint considering heterogeneity of decision makers</td>
<td>Destination choice: Zheng and Guo, 2008</td>
</tr>
</tbody>
</table>
Spatial choice models, including activity location choice models, residential location choice models, migration models etc, therefore include some kind of a choice set generation process. Some spatial choice models (and most residential location choice models) use the random sampling of alternatives to generate choice sets, some models use an importance sampling approach (see for ex. Ben-Akiva and Watanatada, 1981), while others use deterministic or stochastic constraints to limit the choice sets (Scott, 2006, Termansen et al., 2004, Zheng and Guo, 2008). See Table 1 for details.

The methods listed here can be broadly classified into: (a) sampling approaches (e.g. random sampling and importance sampling), and (b) constrained choice set formation models (e.g. deterministic and stochastic constraints). The sampling approaches are purely statistical and are intended to address the issue of computational efficiency. They do not attempt to capture any behavioural realism in the choice set formation process. The constrained choice set formation models, on the other hand, attempt to address the issue of computational efficiency in a more behavioural manner by using deterministic or stochastic conditions to limit the choice set. For instance, a deterministic distance threshold may be used to limit the choice set for activity location, based on the hypothesis that individuals do not travel more than a specified distance. Such a model could be made even more realistic by accounting for inter-personal heterogeneity in the choice set formation process, acknowledging the fact that different individuals have different limiting factors, for e.g. the distance threshold could vary across individuals. However, as the choice set formation model becomes more behaviourally realistic it also loses some computational efficiency. Figure 1 presents a suite of choice set formation models within this context.

<table>
<thead>
<tr>
<th>Behavioural Realism</th>
<th>Computation Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Uniform Random Sampling of Alternatives</td>
<td></td>
</tr>
<tr>
<td>2. Importance Sampling of Alternatives not considering heterogeneity of decision makers</td>
<td></td>
</tr>
<tr>
<td>3. Importance Sampling of Alternatives considering heterogeneity of decision makers</td>
<td></td>
</tr>
<tr>
<td>4. Deterministic constraint not considering heterogeneity of decision makers</td>
<td></td>
</tr>
<tr>
<td>5. Deterministic constraint considering heterogeneity of decision makers</td>
<td></td>
</tr>
<tr>
<td>6. Stochastic constraint considering heterogeneity of decision makers</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 - Behavioural Realism versus Computation Efficiency of different models of choice set formation

In the interest of obtaining operational models, it is therefore necessary to achieve a suitable compromise between behavioural realism, computational efficiency and data needs. This research is aimed at empirically evaluating and comparing the various choice set formation models to identify
the most efficient and behavioural realistic approaches. To the best of our knowledge such an empirical comparison of spatial choice set formation methods has not been attempted before.

In this paper we focus on two-stage spatial choice models, with the choice set formation model implemented in the first stage followed by the spatial choice model in the second stage. This is the most common approach for spatial choice modelling. There are also a few RUM-based choice models that jointly model choice set formation and the choice process (also known as stochastic choice set formation models, for example, the Choice Set Generation Logit (GenL) model by Swait (2001)); however, this is not very common in spatial choice models since the large dimension of the universal choice sets leads to computationally inefficient and often infeasible estimation problems.

The following section describes the modelling framework, with details of the choice set formation models that were empirically tested within a residential location choice context.

### 3. MODELLING FRAMEWORK

In this paper we first implement four different choice set formation models to generate choice sets for each household. This is followed by the estimation of a residential location choice model using each of the choice sets formed with the different models. The base case residential choice model is estimated with the universal choice set for each household. The following choice set formation models are tested in this paper.

1. Random sampling: The choice set for each household is randomly sampled from the universal choice set with a sample size of 50.
2. Deterministic threshold with random sampling: The considered choice set for each households is generated using a deterministic model, and a sample of 50 alternatives is randomly generated from the considered choice set.
3. Importance sampling: The probability of each alternative belonging to the choice set of a household is generated using a model, and a sample of 50 alternatives is derived from the universal choice set using an importance sampling approach.
4. Importance sampling with bias correction: The probability of each alternative belonging to the choice set of a household is generated using a model, and a sample of 50 alternatives is derived from the universal choice set using an importance sampling approach. Further the bias introduced by the importance sampling procedure is corrected in the residential location choice model estimation procedure.

It should be noted that sampling of alternatives from the universal choice set is asymptotically equivalent to using the universal choice set in estimation. In similar manner, random sampling from the considered choice set is asymptotically equivalent to using the entire considered choice sets in estimation. We have also tested this numerically but the results did are not reported here.

In section 3.1, we first present the structure of the residential location choice model that was estimated. The deterministic threshold for choice set formation is modelled in this paper using a hazard-based housing screening model which is described in section 3.2. Each of the choice set formation approaches are then described in section 3.3.
3.1. Residential Location Choice Model Structure

We estimate the residential choice model as a simple Multinomial Logit (MNL) model, which is the dominant approach in most empirical residential location studies. The main reasons which have led to the popularity of MNL models in empirical residential location choice modelling are (a) the computation efficiency of MNL models in terms of estimation and (b) the possibility to estimate MNL models consistently on a subset of choice alternatives.

The MNL residential location choice model generally takes the following form:

\[
V_{z,m} = \sum_{k} \beta_{k}X_{i,k} + \sum_{l} \beta_{l} |X_{i,l} - Y_{n,l}| + \sum_{m} \beta_{m}X_{z,m}Y_{n,m}
\]  

(1)

In the above equation, \(\beta\)'s are parameters to be estimated, \(X\)'s are zonal built-environment variables such as average housing price, and \(Y\)'s are the socio-economic variables of households such as household annual income. It should be noted that there are two forms of interaction terms in the above equation. The former is applicable if the units of variables are the same (ex. zonal average housing price in annualized rent and household annual income), and the latter is applicable if the units of variables are different (ex. zonal household size and household income).

It should be noted that households’ socio-economic attributes cannot be specified by themselves in the utility function because they do not vary across the alternatives and there would be no way to estimate coefficients for such variables. By considering interaction effects between household sociodemographics and the zonal characteristics (As shown in equation 1), households’ attributes are allowed to enter the model in order to take account for the taste variations of households to different locations.

There are some subtle issues to be taken into consideration in the interpretation of model estimation results. On the one hand, if we assume that the universal choice set is the true choice set, then any pruning of the universal choice set will potentially introduce bias. McFadden (1978) demonstrated that the MNL model produces consistent estimates when the choice set is generated using a simple random sampling approach and he also proposed a procedure for correcting the biases that result from non-simple random sampling methods such as importance sampling. On the other hand, if the true choice set is pruned version of the universal choice set (where the pruning reflects the details of the spatial search process) then using the universal choice set or randomly sampled subsets of it, will likewise introduce bias. Since the true choice set is unobserved, this issue is not directly empirically resolvable. Therefore, the principal basis for assessing the alternative approaches to choice set generation is in terms of their prediction performance on a hold-out validation sub-sample (although we do also report on estimated model parameters).

3.2. Hazard-based Housing Screening Model Structure

In this paper, we implement a hazard-based housing screening model as an example of a deterministic choice set formation procedure. Rashidi and Mohammadian (2010) introduced the hazard-based model for housing screening as a behavioural approach to spatial choice set formation.
In the Rashidi and Mohammadian model, the distribution of acceptable housing price and acceptable work commuting distance are modelled using the hazard-based formulation conditional on the household’s socioeconomic attributes. In this section we briefly present the structure of the hazard-based housing screening model.

Let $D$ be a continuous, non-negative valued random variable representing acceptable work commute time beyond which residential alternatives will not be considered by households. And similarly, let $P$ be a continuous, non-negative valued random variable representing maximum affordable housing price for households. Effectively this means that a household will not consider a dwelling or a zone as an alternative if it is more expensive than threshold $P$, or if it is further from the household’s workplace than threshold $D$. This interpretation of acceptable housing price and acceptable commute time is similar to survival analysis where the time it takes for events to occur is examined (time before a failure). For more details of hazard-based models see (Hougaard, 2000).

In this study, we use the Weibull and the Log-logistic distributions for average work commute time and housing price respectively. The results presented in the empirical analysis section justify the use of these distributions. The baseline hazard function and its corresponding survival function for the Weibull model are as follows:

$$\lambda(t) = \alpha \beta t^{\beta-1} \quad S(t) = \exp \left(-\int_0^t \lambda(t) \, dt\right) = \exp \left(-\alpha t^\beta\right) \quad (2)$$

Where, $\alpha$ is the scale parameter and $\beta$ is the shape parameter. In order to incorporate covariates into the model, the scale parameter in the baseline hazard function is reparameterized in terms of explanatory variables and regression parameters and the shape parameter are held fixed (i.e. $\alpha = \exp (\theta_0 - \theta X)$ ). Therefore the proportional hazard function of work commute time and its corresponding survival function are as following:

$$\lambda(t) = \beta t^{\beta-1} \exp (\theta_0 - \theta X) \quad S(t) = e^{-t^\beta \exp (\theta_0 - \theta X)} \quad (3)$$

The baseline hazard function and its corresponding survival function for log-logistic model are as follows:

$$\lambda(t) = \frac{\alpha \beta t^{\beta-1}}{1 + \alpha \beta t^\beta} \quad S(t) = \exp \left(-\int_0^t \lambda(t) \, dt\right) = \frac{1}{1 + \alpha \beta t^\beta} \quad (4)$$

Where, $\alpha$ is the scale parameter and $\beta$ is the shape parameter. Similar to weibull model, the scale parameter in the baseline hazard function is reparameterized in terms of explanatory variables and regression parameters and the shape parameter are held fixed in order to incorporate covariates into the model. Therefore the proportional hazard function of housing price and its corresponding survival function are as follows:

$$\lambda(p) = \frac{\exp (\theta_0 - \theta X) \beta p^{\beta-1}}{1 + \exp (\theta_0 - \theta X) p^\beta} \quad S(p) = \frac{1}{1 + \exp (\theta_0 - \theta X) p^\beta} \quad (5)$$

Based on the definition of hazard function and survival function, the probability density functions of accepting a housing price and accepting a work commute time are as following:

$$f(p) = \lambda(p). S(p), \quad f(t) = \lambda(t). S(t) \quad (6)$$
In order to estimate the model parameters, a joint likelihood function is formulated based on the hazard and survival function across all alternatives’ prices and commute times as follows:

\[ L = \prod_{i=1}^{N} \lambda_p^i \cdot S_p^i, \lambda_t^i \cdot S_t^i \]  

(7)

Where, \( \lambda_p^i \) and \( S_p^i \) are the housing price hazard function and survival function of the household \( i \), and \( \lambda_t^i \) and \( S_t^i \) are the work commute time hazard function and survival function of the household \( i \).

The model parameters include the covariates (Households’ socioeconomic variables such as household income, see empirical analysis section) and Weibull and Log-logistic shape and scale parameters. The joint likelihood function is maximized to estimate the model parameters where housing price and work commute time of households are observed quantities. Then, the probability density functions \( (f(p), f(t)) \) are derived based on estimated parameters for each household.

\[ f_i(t) = \beta t^{\beta - 1} \cdot \exp (\theta_0 - \tilde{\theta} x^i) \cdot e^{-\tilde{\theta} \cdot \exp (\theta_0 - \tilde{\theta} x^i)} \]  

(8)

\[ f_i(p) = \frac{\exp (\theta_0 - \tilde{\theta} x^i) \cdot \beta p^{\beta - 1}}{1 + \exp (\theta_0 - \tilde{\theta} x^i) \cdot \beta p^{\beta - 1}} \cdot \frac{1}{1 + \exp (\theta_0 - \tilde{\theta} x^i) \cdot \beta p^{\beta - 1}} \]  

(9)

### 3.3. Coupling the Choice Set Formation Model with a Discrete Choice Model of Residential Location Choice

As described earlier, each of the choice set formation models is used to generate spatial choice sets for each household, which are then used to estimate residential location choice models. In this section we described how the choice set formation models are coupled with the estimation of the residential choice model.

**Random Sampling**

For each household, the universal choice set is sampled to generate a subset of 50 spatial alternatives. The residential choice model is then estimated on this subset of alternatives. As discussed previously, the MNL model thus estimated is theoretically consistent and unbiased.

**Deterministic Threshold with Random Sampling**

The Rashidi and Mohammadian housing screening model is first implemented and the thresholds on accepting housing price and work commute time are computed for each household from the distribution of acceptable housing prices and work commute times. Here, we assume that household \( i \) considers only zones (spatial choice alternatives) such that the average housing price of that zone is less than the price threshold for the household \( (TP_i) \), and the commute time to the household’s permanent workplace is less than the commute time threshold for the household \( (TT_i) \). To compute the price threshold and commute threshold from the probability density functions of acceptable price and acceptable work commute time we assume that:
which effectively means that zones within the 90th percentile of housing prices and commute times are within the considered choice set for the household. Finally, the choice set for household \( i \) is generated from the considered choice set by making random draws.

This is a commonly used approach in spatial choice models (see for ex. Fotheringham (1988) and Parsons and Hauber (1998)). The choice set for each decision maker in this approach is generated by restricting the choice set to include only the alternatives within a pre-specified threshold for one or more attributes. The sensitivity of parameter estimates to the analyst’s choice of thresholds was examined in the context of recreational destination choice by Parsons and Hauber (1998). They concluded that adding alternatives with the commute time more than three hours in the choice set seem to have negligible effects on the estimation results. The sensitivity of parameter estimates to the choice of thresholds is also examined in this study by estimating different models with different thresholds values.

**Importance Sampling**

Alternately, the choice set for household \( i \) can be generated from the universal choice set using an importance sampling approach. With the preliminary probabilities of selecting alternative \( i \) computed using the housing screening model, different importance sampling strategies can be applied to generate the choice set for a household (Ben-Akiva and Lerman, 1985). In this study, the probability distributions of accepting price and accepting commute time are assumed to be independent; therefore, the joint probability distribution of accepting price and accepting commute time for household \( n \) conditional on its socioeconomic attributes is given by:

\[
f_{n}(p, t) = f_{n}(p) f_{n}(t)
\]  

(11)

Using the results of the hazard-based choice set formation model, the preliminary probability of selecting alternative \( i \) with price \( p_i \) and distance to household \( n \)'s workplace \( d_{i,n} \) can be computed as following:

\[
q_{i,n} = f_{n}(p_i; X_n) f_{n}(t_{i,n}; X_n)
\]  

(12)

In order to generate the estimation choice set of household \( n \), we define a sampling protocol as follows. A set \( C_n \) is generated by drawing \( N \) (Sample Size = 50 in this study) alternatives with replacement from the universal set, and then adding the chosen alternative to it. Each alternative has a sampling weight of \( q_{j,n} \). The outcome of this protocol is \((k_{1,n}, k_{2,n}, ..., k_{j,n})\), where \( k_{j,n} \) is the number of times alternative \( j \) is drawn.

**Importance Sampling with Bias Correction**

As discussed by McFadden (1978), the MNL estimated using a non-random sample of alternatives is biased and needs an alternative-specific correction term to adjust for the bias. In this study we use
an alternative-specific correction term for this sampling protocol as derived by Freijinger et al. (2009).

\[ \pi_i(C_n|t) = K_{C_n} \frac{k_{i,n}}{q_{i,n}} \]  

(13)

where \( k_{i,n} \) and \( q_{i,n} \) are as defined in the previous section.

It should be noted that the \( K_{C_n} \) cancels out since it is constant for all alternative in \( C_n \) and the final ‘corrected’ probability that household \( n \) chooses alternative \( i \) in \( C_n \) is:

\[ P(i|C_n) = \frac{e^{\theta v_{i,n} + \ln \left( \frac{k_{i,n}}{q_{i,n}} \right)}}{\sum_{j \in C_n} e^{\theta v_{j,n} + \ln \left( \frac{k_{j,n}}{q_{j,n}} \right)}} \]  

(14)

4. EMPIRICAL ANALYSIS

In this section, we present the results of the empirical analysis. Section 4.1 describes the data sources in relation to the desired specification of the residential location choice model, section 4.2 describes data preparation specifically for the hazard-based housing screening model, section 4.3 presents the model estimation results and section 4.4 presents the validation results and a discussion of the results.

4.1. Data Sources and Variable Specification for the Residential Choice Model

The area selected for this study is the Greater London comprising the 32 London boroughs and the City of London. In 2001 London had 7,172,091 inhabitants and 3,015,997 households (Census 2001). The Greater London covers 1,594.72 square kilometres which occupies 1.2 percent of England.

The primary data source used in this analysis is the London Area Transport Survey (LATS). The estimation dataset used for this study is a subset of 12836 households from the LATS 2001 data, a subset generated by cleaning the data, and considering only households within the Greater London administrative boundaries that have at least one worker. The spatial choice alternatives for residential location are LTS (London Transport Studies) model zones which are called Transport Analysis Zones (TAZ) in this study (For more details, refer to ‘zn02’ LTS zoning system in LTS documentation, Transport for London (TfL)).

4.1.1. Built-environment Variables

In addition to the LATS data, several other data sets were used to obtain spatial and built-environment variables for the study area. The data sources uses in this study are as following.

1. Generalized Land-Use Database 2001
2. Census 2001
3. Zone-to-zone travel time of auto mode
4. Housing price data from Land registry

Since the TAZs are assumed to be residential alternatives, all variables are matched and aggregated to TAZ level.
These spatial data sources provide a rich set of variables for consideration in the model specification. These variables can be broadly categorized into 10 groups including zonal size and density, zonal ethnic composition, zonal land use structure, zonal demographics and housing price, regional accessibility measures, local transport network measures, activity opportunities and commute related variables. The specific formulation used for the zonal land use structure, regional accessibility measures and local transport network measures are described below.

**Zonal land-use structure**
These variables relate to land-use composition of zones and have been derived from Generalized Land-Use Database (GLUD) 2001. These data are available at the census output area (OA) level and have been aggregated to TAZs. Diversity of land-uses and the degree to which they are mixed might be a determinant zonal attributes in residential location choice. Guo and Bhat (2007) used land-use mix diversity index in order to be able to measure the zonal land-use mix and incorporate it in a residential location choice model. This measure aims to capture the mix of uses relative to a perfect distribution of uses. Similar to Guo and Bhat (2007), the land-use mix diversity index in this study includes three different land uses as shown below.

\[
Land-use\, mix\, diversity = 1 - \left( \frac{\left| \frac{d}{L} - \frac{1}{3} \right| + \left| \frac{n}{L} - \frac{1}{3} \right| + \left| \frac{o}{L} - \frac{1}{3} \right|}{\frac{4}{3}} \right)
\]

(15)

where \( L = d + n + o \), \( d \) is zonal area in domestic building use (residential), \( n \) is zonal area in non-domestic building (non-residential) use and \( o \) is zonal area in other uses.

**Regional accessibility measure**
In this research, accessibility of zones to employment, shopping and recreational opportunities have been computed using annual business inquiry data. Four Hansen-type accessibility measures are developed as follows:

\[
A_{i, \text{Rec}}^{\text{Rec}} = \frac{1}{N} \sum_{j=1}^{N} \frac{R_j}{TT_{ij}} \quad A_{i, \text{Shop}}^{\text{Shop}} = \frac{1}{N} \sum_{j=1}^{N} \frac{S_j}{TT_{ij}} \quad A_{i, \text{Health}}^{\text{Health}} = \frac{1}{N} \sum_{j=1}^{N} \frac{H_j}{TT_{ij}} \quad A_{i, \text{Emp}}^{\text{Emp}} = \frac{1}{N} \sum_{j=1}^{N} \frac{E_j}{TT_{ij}}
\]

(16)

where \( R_j \) is number of employees in recreational, cultural and sporting activities sector, \( S_j \) is number of employees in retail trade sector, \( H_j \) is number of employees in health and social work sector, \( E_j \) is total number of employees and \( TT_{ij} \) is the travel time from zone \( i \) to zone \( j \).

**Local transportation network measure**
The local transportation network measures are included because they represent local measures of public transportation and auto levels of service of a zone. OS MasterMap ITN layer has been used to compute Dual carriageway density and other roadway density variables.

Tube, DLR and Overground serve a large part of Greater London and neighbouring areas of Essex, Hertfordshire and Buckinghamshire. They have accessibility and environmental impacts which can
affect households’ residential location choice decisions. In this study, transit accessibility measure of a zone is defined as the summation of inverse distances from all stations to the centroid of the zone.

\[ A_i^{\text{transit}} = \sum_{j=1}^{N} \frac{1}{d_{ij}} \]  

(17)

where \(d_{ij}\) is the distance of zone \(i\) to station \(j\), and \(N\) is the total number of stations.

4.1.2. Interaction of household sociodemographics with zonal characteristics

In addition to the variables described in the previous section, several interaction terms are also considered in the model:

1. Interaction of annual household income and average housing price in annualized rents which represents household disposable income
2. Interaction of household size with average zonal household size
3. Interaction of household income with zonal density
4. Interaction of household income with accessibility measures (accessibility to employment)
5. Interaction of household income with commute variables

4.2. Specification of the Hazard-based Choice Set Formation Model

4.2.1. Dependent Variables

As mentioned earlier, two dependent variables were used to model the screening process of residential locations -- average work commute time and average housing price. Average work commute time as aggregate values across all workers in the household can be computed from LATS 2001. Average housing prices data can also be computed from Land registry disaggregate data set. Kolmogorov-Smirnov test (K-S test) of different distributions confirms that the average work commute time follows a Weibull distribution, and average housing prices follows a Log-logistic distribution (see Figures 2, 3 and Table 2).

![Figure 2 - Average Work Commute Time Distribution](image)
Table 1 - Kolmogorov-Smirnov Statistics of Different Distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Average Work Commute Time</th>
<th>Average Housing Price</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K- S Statistic</td>
<td>Rank</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.0862</td>
<td>4</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.1096</td>
<td>5</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>0.0559</td>
<td>2</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.0823</td>
<td>3</td>
</tr>
<tr>
<td>Normal</td>
<td>0.1699</td>
<td>6</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.0323</td>
<td>1</td>
</tr>
</tbody>
</table>

4.2.2. Explanatory Variables
Households’ socioeconomic variables such as household income, household size, number of workers in the household, number children in household and households’ vehicle ownership are behaviourally considered as influential factors in determining probability distribution of acceptable housing price and work commute time for a household. Therefore, they have been included in the model.

4.3. Estimation Results

4.3.1. Hazard-based Choice Set Formation Model
The results of parameter estimation of the hazard-based choice set formation model are tabulated in table 2.
Table 2 – Summary Statistics of Expected Choice Set

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>STD</th>
<th>t – statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acceptable Housing Price (Log-logistic distribution)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No 1 Theta (Constant)</td>
<td>-26.7808</td>
<td>0.2032</td>
<td>-131.787</td>
</tr>
<tr>
<td>No 2 Beta (Shape parameter)</td>
<td>5.1389</td>
<td>0.0385</td>
<td>133.4298</td>
</tr>
<tr>
<td>No 3 Number of household members</td>
<td>-0.2316</td>
<td>0.0126</td>
<td>-18.2731</td>
</tr>
<tr>
<td>No 4 Number of children age 1-5</td>
<td>0.1469</td>
<td>0.0353</td>
<td>4.1514</td>
</tr>
<tr>
<td>No 5 Average household income (x1000)</td>
<td>0.0162</td>
<td>0.0007</td>
<td>22.8498</td>
</tr>
<tr>
<td><strong>Acceptable Commute Time to Work (Weibull distribution)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No 1 Theta (Constant)</td>
<td>-4.4902</td>
<td>0.0435</td>
<td>-103.222</td>
</tr>
<tr>
<td>No 2 Beta (Shape parameter)</td>
<td>1.2562</td>
<td>0.0091</td>
<td>139.4595</td>
</tr>
<tr>
<td>No 3 Number of vehicles</td>
<td>-0.0346</td>
<td>0.0108</td>
<td>-3.1928</td>
</tr>
<tr>
<td>No 4 Number of workers</td>
<td>-0.1009</td>
<td>0.0134</td>
<td>-7.4869</td>
</tr>
<tr>
<td>No 5 Average household income (x1000)</td>
<td>0.0099</td>
<td>0.0004</td>
<td>22.7236</td>
</tr>
</tbody>
</table>

**Summary statistics**
- Number of Observations: 12836
- Value of the log-likelihood function at zero: -198824
- Value of the log-likelihood function at convergence: -129485
- -2(L(0)-L(β)): 138678

It should be noted that the effect of covariates in a hazard model is facilitated by incorporating a negative sign for the parameters in the formulation. In other words, if a covariate gets a negative sign, the chance of failure, or equivalently the probability of accepting a price or a work distance, is increased. Alternately, a positive sign suggests that any increase in the covariate decreases the chance of failure for the household which implies that the household tends to pay more or travel longer.

The negative sign for the number of household members in the model suggests that in the screening process large households are looking for less expensive areas. This can be justified by the fact that large households presumably need more space. Therefore, all else being equal, they tend to obtain more housing space for their limited budget in less expensive neighbourhoods. On the other hand, the positive sign of number of children aged 1-5 suggests that households with small children prefer more expensive neighbourhoods, which could be a proxy for safety and school quality. Furthermore, the positive sign of the average household income shows the obvious fact that wealthier households tend to live in more expensive neighbourhoods.

Household income which is positively correlated with the number of vehicles is also important on the household decision about the average work commute time. The positive signs of household income and number of vehicles show that the wealthier households are also more likely to live in
suburban areas and commute farther to their workplaces. On the other hand, negative sign of number of workers suggests that households with more workers tend to live closer to their workplace. This can be justified by the fact that work commute time in residential location screening process is more important for households with more workers since they need to spend more travel cost and time in total if they want to live further away comparing with the households with fewer workers.

In general, the hazard-based choice set formation model appears to yield intuitive and behaviourally realistic results.

4.3.2. Residential Location Choice Model

The built environment variables, commute-related variables and interaction variables introduced in the previous section were identified from the literature as likely determinants of a household’s choice of residential location.

Several specifications of the residential choice model have been tested and statistically insignificant variables have been systematically eliminated in order to find the true model for residential location choice of Greater London. In the final model specification, the following variables have been considered.

- Log of zonal area (Hectares)
- Number of residents per Hectares
- Average household size
- Percentage of zonal area occupied by Domestic Buildings
- Housing Price 1/1000
- Accessibility to Shopping
- Accessibility to Recreational facilities
- Accessibility to Employment
- Absolute difference of household income and annualized rent
- Absolute difference of household size and average household size
- Travel time from households’ work zone to candidate residential zones

In order to examine the performance of the different approaches, five models of residential location have been estimated with different assumptions regarding the household’s choice sets.

1. UCS(Universal Choice Set): Households’ choice set is the universal choice set
2. SRS (Simple Random Sampling): Households’ choice set is randomly sampled (uniform sampling) form the universal choice set
3. DC (Deterministic Constraint): Households’ choice set is generated using deterministic constraints computed from the screening model
4. ISC (Importance Sampling adding Correction terms): Households’ choice set is generated using importance sampling of alternatives based on prior probabilities of choosing alternatives computed from the screening model and adding alternative-specific correction terms in order to correct for bias of importance sampling
5. ISNC (Importance Sampling Not adding Correction terms) Households’ choice set is generated using importance sampling of alternatives based on prior probabilities of choosing
alternatives computed from the hazard-based model and **NOT** including the alternative bias correction terms.

Based on the result of the housing screening model, the expected choice set can be generated for each household conditional on the price and distance thresholds of each household. Since the housing price and the commute time thresholds are different for each household depending on their socioeconomic variables, the size of the expected choice sets are different too. The mean size of the expected choice sets is 389 (The universal choice set contains 861 alternatives). This indicates that the housing screening model is capable of reducing the size of universal choice set by over 50%.

The estimation results of five above mentioned residential choice models are presented in Table 3. In model estimation, the full dataset (100 percent) have been used. In general, the coefficient of the log of the zonal area has a positive sign as expected, indicating that households are more likely to locate themselves in zones with a large number of housing opportunities. This parameter is significant in all models. The coefficient of population density (number of residents per hectare) is also positive and significant in all models. The zonal average household size and its interaction with household size are negative and significant in all models. This confirms the clustering of households based on the zonal household size observed in previous studies.

Intuitively, accessibility to employment has a positive effect on the utility but the coefficient of the accessibility to employment variable is found to be negative here. Similarly Guo and Bhat (2004) found a negative utility of accessibility to work for African-American households in the Dallas county area. The residential choice models estimated in this study likewise suggest that most people in London live in neighbourhoods that are less accessible to employment opportunities. This is a reflection of the housing market in London, the most desirable neighbourhoods typically being located far from the major employment centres. On the other hand, the positive sign of the accessibility to shopping variable indicates that households prefer locations that offer good accessibility to shopping.

The commute-related variables are important determinants of residential choice. Households, in general locate to reduce their commute time. Therefore, commute time is expected to have a negative sign, which is the case in all models.

Zonal housing price and its interaction with household income is expected to have a negative effect on the utility of a residential choice alternative, which is also addressed in previous studies. This parameter was found to be negative in UCS, SRS and ISC. But, interestingly, the zonal housing price parameter and its interaction with households’ average income are positive in DC and ISNC. This result is unexpected and initially somewhat counter-intuitive. However, it may reflect a selectivity effect in the operation of the screening model. In particular, it could be argued that the use of a price threshold in the screening model has the effect of essentially filtering out of the choice set unaffordable locations, leaving behind those that are affordable. Amongst these affordable locations, price is likely to be positively correlated with unobserved quality attributes, implying that households in effect choose the best location that they can afford amongst the screened alternatives.
<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>1. UCS (Universal Choice Set)</th>
<th>2. SRS (Simple Random Sampling)</th>
<th>3. DC (Deterministic Constraint)</th>
<th>4. ISC (Importance Sampling adding correction terms)</th>
<th>5. ISNC (Importance Sampling NOT adding Correction terms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimate</td>
<td>t-statistic</td>
<td>Estimate</td>
<td>t-statistic</td>
<td>Estimate</td>
</tr>
<tr>
<td>1</td>
<td>Log of zonal area (Hectares)</td>
<td>2.2999</td>
<td>50.72</td>
<td>2.2904</td>
<td>48.4215</td>
<td>2.3376</td>
</tr>
<tr>
<td>2</td>
<td>Number of residents per Hectares</td>
<td>0.0115</td>
<td>32.30</td>
<td>0.0107</td>
<td>26.0512</td>
<td>0.0116</td>
</tr>
<tr>
<td>3</td>
<td>Average household size</td>
<td>-0.4413</td>
<td>-11.08</td>
<td>-0.4107</td>
<td>-9.577</td>
<td>-0.4238</td>
</tr>
<tr>
<td>4</td>
<td>Percentage of zonal area occupied by Domestic Buildings</td>
<td>0.6902</td>
<td>4.88</td>
<td>0.596</td>
<td>4.0448</td>
<td>0.6459</td>
</tr>
<tr>
<td>5</td>
<td>Housing Price 1/1000</td>
<td>-0.0025</td>
<td>-12.52</td>
<td>-0.0024</td>
<td>-11.6376</td>
<td>-0.0027</td>
</tr>
<tr>
<td>6</td>
<td>Accessibility to Shopping</td>
<td>0.1402</td>
<td>22.86</td>
<td>0.1317</td>
<td>21.3869</td>
<td>0.1467</td>
</tr>
<tr>
<td>7</td>
<td>Accessibility to Recreational facilities</td>
<td>-0.1396</td>
<td>-7.04</td>
<td>-0.1402</td>
<td>-6.973</td>
<td>-0.1609</td>
</tr>
<tr>
<td>8</td>
<td>Accessibility to Employment</td>
<td>-0.0229</td>
<td>-19.00</td>
<td>-0.0212</td>
<td>-17.4308</td>
<td>-0.0224</td>
</tr>
<tr>
<td>9</td>
<td>Absolute difference of household income and annualized rent</td>
<td>-0.0074</td>
<td>-3.38</td>
<td>-0.0087</td>
<td>-3.8428</td>
<td>-0.0106</td>
</tr>
<tr>
<td>10</td>
<td>Absolute difference of household size and average household size</td>
<td>-0.6762</td>
<td>-18.35</td>
<td>-0.6873</td>
<td>-17.431</td>
<td>-0.6861</td>
</tr>
<tr>
<td>11</td>
<td>Travel time from households’ work zone to candidate residential zones</td>
<td>-0.0619</td>
<td>121.73</td>
<td>-0.0596</td>
<td>115.4077</td>
<td>-0.0627</td>
</tr>
</tbody>
</table>

**Summary statistics**

- **Number of Observations**: 12836
- **Sample Size (Choice set size)**: 861
- **Value of the log-likelihood function at zero**: 86740.1
- **Value of the log-likelihood function at convergence**: 72673.4
- **\( \rho^2 \)**: 0.1622

---

**Table 3 – Models Estimation Results**

---
4.4. Sensitivity of Parameter Estimates

In the estimation of DC model we assumed that an alternative is considered by a household if and only if the average housing price of that alternative is less than the household’s price threshold and the commute time to the household’s permanent workplace is less than the household’s commute time threshold. Further we assumed that the household’s price threshold and commute time threshold are the values from probability distributions of accepting housing price and accepting commute time corresponding to the probability of 0.9 (90th percentile). As mentioned earlier, the choice of the percentile depends on the analyst’s judgement. In this section, we examine the sensitivity of the parameter estimates of the residential location choice model to different threshold values.

It is obvious that the choice set boundary expands as the corresponding percentile of housing price and commute time thresholds increase i.e. the constraints on price and commute time are relaxed. Hence, the mean size of the choice set increases as the thresholds are relaxed, as tabulated in table 4.

<table>
<thead>
<tr>
<th>Boundary Definition (Percentile)</th>
<th>Mean Number of Alternatives in a Household’s Choice Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>195.23</td>
</tr>
<tr>
<td>0.85</td>
<td>271.13</td>
</tr>
<tr>
<td>0.9</td>
<td>388.75</td>
</tr>
<tr>
<td>0.95</td>
<td>580</td>
</tr>
<tr>
<td>0.99</td>
<td>811.14</td>
</tr>
</tbody>
</table>

With the exception of commute time and housing price coefficients (as well as interaction of housing price with annualized rent), all the coefficients in residential location choice model show little change as the boundary expands. But the housing price and commute time coefficients change dramatically as the value of the percentile (therefore housing price and commute time thresholds) changes. This has been depicted in Figures 4 and 5.
Figure 4 – Housing Price Coefficient Estimates

Figure 5 – Workplace Commute Time Coefficient Estimates
The commute time coefficient has the expected negative sign in all the models, however this coefficient becomes more negative (implying a greater disutility attributed to compute time) when the boundary expands. This shows that when the effect of commute time is modelled as a non-compensatory process in the choice set formation stage (by assuming commute time thresholds), the commute time coefficient is estimated to have less impact in the compensatory stage when compared to a choice model that considers the universal choice set. In other words, incorporating a constraint on the commute time (non-compensatory process) in the choice set formation stage fades the effect of commute time in the compensatory stage and vice versa.

One can argue that the choice set formation stage is only a reflection of households’ preferences regarding commute time. Horowitz and Louviere (1995), for instance, argue that modelling choice behaviour as a two-stage process (choice set formation stage followed by actual choice from the smaller choice set) provides no information beyond that contained in the utility function. Hence, the choice set formation stage is simply an indicator of preferences. Incorporating a choice set formation stage in choice modelling has largely been justified on the grounds that including a choice set formation stage (a) provides a more realistic representation of the choice process and (b) leads to improved forecasts and better explanation of behaviour. However, Horowitz and Louviere (1995) show that, within their empirical setting, modelling choice as a two-stage process may lead to a mis-specified model that makes erroneous forecasts. Moreover, in the absence of direct observation of choice processes we cannot know that the two stage nature of decision making is the most behaviourally realistic.

In our empirical setting, as illustrated in figures 4 and 5, the rho square (goodness of fit) of the models increases as the boundary expands. This effectively means that a model with the full choice set is a better fit for the data in terms of the value of likelihood ratio index. Models’ prediction performances in terms aggregate and disaggregate prediction performance have discussed in the next section.
4.5. Model Validations and Discussion

The model estimation results indicate that assuming the universal choice set (UCS) for households and applying sampling strategies, i.e. uniform random sampling (SRS) and importance sampling with correction (ISC), are equivalent. This agrees with the theory, which states that sampling of alternatives is a statistical solution to cope with large numbers of alternatives and is asymptotically equivalent to using the universal choice set. In fact, UCS, SRS and ISC represent the same assumed behaviour, which is that the universal choice set is the actual choice set considered by households.

On the other hand, the application of deterministic constraints to filter the alternatives and generate considered choice sets (DC) is hypothesised to represent a different behaviour. As discussed by Rashidi and Mohammadian (2010), this approach attempts to model choice set formation behaviour in a more realistic manner under the hypothesis that there are two thresholds for each household (housing price and work commute time) that decide the considered choice set.

One of the findings of this study is that the estimated parameters of DC (deterministic constraint) and ISNC (importance sampling not adding correction terms) are similar. In fact, it can be theoretically argued that these two approaches of choice set generation are equivalent. The deterministic choice set formation approach can be viewed as an importance sampling protocol where alternatives are assigned a weight of zero if they are outside the thresholds and a weight of 1 if they are inside the thresholds. Looked at from this perspective, the deterministic choice set formation model must also yield biased parameters (like the importance sampling approach) unless appropriately corrected. However, since the deterministic model is hypothesised to be a behavioural model of housing screening, and therefore choice set formation, the resulting choice sets are assumed to be the true choice sets considered by the households and therefore require no bias correction.

In the validation of the models, model estimations were achieved using a randomly drawn 75% of the data, and the remaining 25% were kept aside as a hold-out sample. The 25% hold-out sample is used to compare the performance of the choice set formation models. The performance of the models was compared using different measures: (a) disaggregate validation and (b) aggregate validation.

Similar to Bhat and Pulugurta (1991), for disaggregate validation, two measures of fit have been used. The first measure is the predictive adjusted likelihood ratio index. This measure is computed by calculating the predictive log-likelihood function value at the parameter estimates obtained by maximizing the estimation likelihood function and then computing the corresponding adjusted likelihood ratio index. The second measure is the average probability of correct prediction which has been computed for different models as following:

\[ N^{-1} \sum \sum \gamma_{ij} \bar{p}_{ij} \]

where \( N \) is the number of observations in the validation sample, \( \gamma_{ij} \) is a dummy variable indicating if household \( i \) lives in zone \( j \), and \( \bar{p}_{ij} \) is the predicted probability of household \( i \) is located in zone \( j \).

For aggregate validation, the predicted and actual market shares of the hold-out sample have been computed for different models. The root mean square errors (RMSE) and the mean absolute
deviation (MAD) of the predicted shares have been used to compare the performance of different models.

Table 5-Prediction Test on the Hold-out Sample

<table>
<thead>
<tr>
<th>Summary Statistic</th>
<th>UCS</th>
<th>SRS</th>
<th>DC</th>
<th>ISC</th>
<th>ISNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value of the log-likelihood function at zero (Validation Sample)</td>
<td>21687</td>
<td>12554</td>
<td>12535</td>
<td>13795</td>
<td>12419</td>
</tr>
<tr>
<td>Value of the predicted log-likelihood function¹</td>
<td>18053</td>
<td>9124</td>
<td>10283</td>
<td>10288</td>
<td>10521</td>
</tr>
<tr>
<td>Predicted Likelihood ratio index</td>
<td>0.168</td>
<td>0.273</td>
<td>0.180</td>
<td>0.254</td>
<td>0.153</td>
</tr>
<tr>
<td>Average Probability of Correct Prediction</td>
<td>0.0011</td>
<td>0.0178</td>
<td>0.0198</td>
<td>0.0196</td>
<td>0.0194</td>
</tr>
<tr>
<td>RMSE²</td>
<td>2.851</td>
<td>2.577</td>
<td>3.351</td>
<td>3.611</td>
<td>2.867</td>
</tr>
<tr>
<td>MAD³</td>
<td>2.089</td>
<td>1.926</td>
<td>2.361</td>
<td>2.218</td>
<td>2.092</td>
</tr>
</tbody>
</table>

¹ The predictive log-likelihood is the log-likelihood function value in the validation sample computed at the parameter estimates obtained from maximizing the estimation likelihood function

² RMSE represents the root mean square error between predicted and actual shares

³ MAD represents the mean absolute deviation between predicted and actual shares

Here, UCS, SRS and ISC are statistically equivalent, and as mentioned earlier, DC and ISNC are equivalent as well. Hence, in validation of models we expect to achieve similar results for equivalent models. It should be noted that results of UCS should not be compared to others because the choice probabilities for UCS model have been computed over a much larger choice set (Universal Choice Set).

The estimated parameters of models also confirm that UCS, SRS and ISC are equivalent and DC and ISNC are equivalent, but the validation results present a somewhat mixed picture. The aggregate RMSE and MAD statistics and the disaggregate predicted likelihood ratio index criteria suggest that SRS is the superior approach, while the average probability of correct prediction criterion suggests that DC performs slightly better than SRS. Hence, on balance, the validation results seem to be in favour of SRS model which is asymptotically equivalent to the UCS model.

The RMSE and MAD statistics as well as average probability of correct prediction criterion have also been computed for different values percentile of DC model as described in the previous section. The results confirm that as the boundary expands the RMSE and MAD decreases which effectively mean that the model with the full choice set and using simple random sampling strategy is the superior model. With respect to the average probability of correct prediction criterion, here once again the results present a mixed picture since as the boundary expands this measure decreases which means
that the model with full choice set and using simple random sampling strategy is the inferior model.
In conclusion, the results present no clear evidence that DC is effective in producing superior
predictive performance.

5. Conclusion
In this paper, we have compared the empirical performance of a recently developed hazard-based
model of housing search choice set formation (Rashidi and Mohammadian, 2010) with more
conventional statistical methods of choice set pruning. The principal basis for the comparison is in
terms of their prediction performances on a hold-out validation sub-sample. Despite its intuitive
appealing behavioural plausibility, the results of aggregate and disaggregate validation of models
suggest that the hazard based choice formation model in fact performs worse than the alternative
statistical pruning approaches.

To date most effort in the residential location literature has been directed at improving the
specification of the choice process by for example, accommodating complex patterns of unobserved
spatial correlation amongst alternatives. The finding of this study emphasises both the importance
of modelling choice set formation and the high level of challenge involved in doing so effectively in
the context of residential location choice. It must be remembered that any choice set formation
model is attempting to characterise what is an underlying highly complex and dynamic process of
housing market search, which will depend on the spatial dynamics of labour markets and housing
supply, market intermediator (e.g., estate agent) activity, mortgage and interest policy as well as
many perceptual and cognitive factors. A simple threshold model based on housing cost and
commute time alone is unlikely to capture well this complex set of influences, and any systematic
mis-specification in the choice set formation model will be propagated into the resulting choice
model, and reflected in a degraded prediction performance (as well as impacts on estimation of key
model parameters, such as the importance of house price).

There is an urgent need for researchers to identify data sources that can usefully inform on the
underlying drivers of choice set formation in the residential location market and to integrate such
data into operationally tractable models that offer genuine improvements in predictive
performance.
References


MCFADDEN, D. 1978. Modelling the choice of residential location, Institute of Transportation Studies, University of California.


