The Journal of Transport and Land Use http://jtlu.org
Vol. 12 No. 1 [2019] pp. 597–618

JTLU

Modelling residential location choices with implicit availability of alternatives

Md Bashirul Haque
University of Leeds
tsmbh@leeds.ac.uk
and
Shahjalal University of Science and Technology
bashir-cee@sust.edu

Charisma Choudhury
University of Leeds
C.F.Choudhury@leeds.ac.uk

Stephane Hess
University of Leeds
s.hess@its.leeds.ac.uk

Abstract: Choice set generation is a challenging aspect of disaggregate level residential location choice modelling due to the large number of candidate alternatives in the universal choice set (hundreds to hundreds of thousands). The classical Manski method (Manski, 1977) is infeasible here because of the explosion of the number of possible choice sets with the increase in the number of alternatives. Several alternative approaches have been proposed in recent years to deal with this issue, but these have limitations alongside strengths. For example, the Constrained Multinomial Logit (CMNL) model (Martínez et al., 2009) offers gains in efficiency and improvements in model fit but has weaknesses in terms of replicating the Manski model parameters. The rth-order Constrained Multinomial Logit (rCMNL) model (Paleti, 2015) performs better than the CMNL model in producing results consistent with the Manski model, but the benefits disappear when the number of alternatives in the universal choice set increases. In this study, we propose an improved CMNL model (referred to as Improved Constrained Multinomial Logit Model, ICMNL) with a higher order formulation of the CMNL penalty term that does not depend on the number of alternatives in the choice set. Therefore, it is expected to result in better model fit compared to the CMNL and the rCMNL model in cases with large universal choice sets. The performance of the ICMNL model against the CMNL and the rCMNL model is evaluated in an empirical study of residential location choices of households living in the Greater London Area. Zone level models are estimated for residential ownership and renting decisions where the number of alternatives in the universal choice set is 498 in each case. The performance of the models is examined both on the estimation sample and the holdout sample used for validation. The results of both ownership and renting models indicate that the ICMNL model performs considerably better compared to the CMNL and the rCMNL model for both the estimation and validation samples. The ICMNL model can thus help transport and urban planners in developing better prediction tools.

Keywords: Choice set generation, Manski model, constrained multinomial logit model, Greater London

Article history:
Received: August 16, 2018
Received in revised form:
January 14, 2019
Accepted: March 3, 2019
Available online: July 24, 2019

http://dx.doi.org/10.5198/jtlu.2019.1450
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

1 Introduction

Home location and urban environment determine the activities and travel patterns of individual household members. For example, households living in inner-city areas with closer proximity to facilities tend to travel shorter distances, make more non-motorized trips and are found to be less car-dependent for local transport than suburbanites (Næss, 2009). These prevailing dynamics and interactions among residential location, urban form and travel behaviour are at the heart of integrated land use and transport planning. Therefore, modelling residential location choice is a key component of integrated planning.

Modelling of residential location choices has numerous challenges from both a methodological and empirical standpoint. A key issue in this regard is choice set formulation, which has a substantial effect on the model outputs (Swait, 2001; Bell, 2007). In a standard choice model, an analyst needs to specify all the alternatives considered by the decision makers. However, in the context of residential location choice modelling, typically the individual-level choice set is unknown to the analyst and the universal choice set is very large – hundreds in case of zone level to hundreds of thousands in the case of dwelling level models. A review of the literature reveals several residential location choice models where the universal choice set has been used as the individual choice set (e.g., Bhat & Guo, 2004; Zolfaghari, 2013; Haque, Choudhury, & Hess, 2018).

Considering the universal choice set for individuals is however also behaviourally unrealistic, as in the real world, households are neither aware of the full set of alternatives nor consider all alternatives they are aware of. Different households might thus have different consideration sets based on household preferences and sociodemographic characteristics, as well as their knowledge of available alternatives. For example, a household may not consider an alternative if they do not have enough knowledge about it or if the alternative is very far from the workplace of a household member. Therefore, it is expected that better ways to model the choice set will make the models behaviourally more representative. This, in turn, will lead to more accurate models for planning and policy making.

Modelling of individual choice sets from a large universal choice set (e.g., disaggregate level residential location choice modelling) using the two-stage probabilistic approach (e.g., Manski 1977; Swait & Ben-Akiva, 1987, etc.) is infeasible due to the explosion of estimation complexity with increasing numbers of alternatives in the universal choice set. Therefore, a two-stage deterministic approach has been used in several studies where a limited set of alternatives have been screened from a universal choice set for each individual based on some behavioural rules (e.g., Farooq & Miller, 2012; Rashidi, Auld, & Mohammadian, 2012; Zolfaghari, 2013, etc.). However, this technique has a high risk of excluding potentially viable alternatives from the individual choice set and including irrelevant alternatives. The performance of the deterministic choice set generation approaches has been evaluated and criticized in the literature (Zolfaghari, 2013). A single-stage semi-compensatory technique (e.g., Cascetta & Papola, 2001, 2009: Murtinez, Aguila, & Hurtubia, 2009; Paleti, 2015) provides an alternative approach to avoiding the risk of screening of alternatives in the two-stage deterministic approaches. This approach captures the individual choice set implicitly through a form of utility penalization and is computationally feasible in the case of large universal choice sets. However, semi-compensatory methods also have limitations in their ability to reproduce the true parameters in estimation (Bierlaire, Hurtubia, & Flötteröd, 2010).

Based on the above discussion, all existing methods have weaknesses alongside strengths. In particular, it is unclear if a specific approach is better for a particular choice context (e.g., residential location vs. route choice). Furthermore, previous work has either focused only on choice set generation for long-term residential location choices (e.g., ownership) or did not make any distinction between the long and medium term (e.g., renting) decisions. Our own previous research (Haque et al., 2018) has however indicated significant differences between the sensitivities to different parameters in the two different residential choice contexts.

Motivated by these points, the specific objectives of this study are as follows:

- to evaluate the performance of state-of-the-art semi-compensatory choice set generation techniques in the context of residential location choice modelling;
- to investigate the potential to improve choice set generation techniques without compromising computational tractability; and
- to investigate the existence of underlying heterogeneity in the choice set of long-term and medium-term residential location choices (residential ownership and renting respectively).

In the next two sections, we review the choice set generation techniques in further detail and propose an improved choice set generation technique. The data for our empirical example is presented next followed by the estimation and the validation results. We conclude with a summary of the findings and directions for future research.

2 Review of choice set generation techniques

The two-stage probabilistic approach proposed by Manski (1977) is a classical solution for modelling individual choice sets. This method requires the estimation of probabilities of all possible choice sets in the first stage and the conditional probabilities of alternatives across all choice sets in the second stage. Both stages are estimated simultaneously. The number of possible choice sets explodes with the number of alternatives in the universal choice set. For J alternatives, the number of possible choice sets is 2J-1. Therefore, this method is computationally infeasible for a medium to large choice set (e.g., residential location choice, route choice, destination choice, etc.).

The unconditional probability of choosing an alternative by a decision maker in the Manski method is the product of the conditional probability of the alternative (given the choice set) and the probability of the choice set. The probability can be presented as follows:

$$P_{in}(C)=\Sigma_{C_s \in C} P_n(i/C_s)\times P(C_s) \quad (1)$$

where $P_{in}(C)$ is the unconditional probability of choosing alternative i by individual n from universal set C, $P_n(i/C_s)$ is the conditional probability of choosing alternative i from the choice set C_s ($C_s \in C$) and $P(C_s)$ is the probability of the choice set being C_s. We thus have a sum across all the possible choice sets.

Other probabilistic approaches proposed in the literature as alternatives to the Manski method (e.g., Swait & Ben-Akiva, 1987; Swait, 2001; Kaplan, Bekhor, & Shiftan, 2011; Zolfaghari, 2013; Bhat, 2015, etc.) also have computational issues for large numbers of alternatives.

The typical approach in the literature for modelling with large universal choice sets is random selection of a subset of alternatives from the universal choice set (Bhat & Guo, 2007; Habib & Miller, 2009; Lee & Waddell, 2010; Guevara, 2010). This approach reduces the computational burden substantially and can also produce consistent parameters in estimation if households consider all alternatives in the universal choice set (McFadden, 1978). This is, however, a poor assumption in the case of disaggregate level residential location choice modelling. Rather, it seems more reasonable that households apply behavioural process heuristics for screening of the alternatives (Bhat, 2015). Ignoring this underlying search mechanism can potentially lead to inaccurate parameter estimation, wrong model forecasts and inappropriate policy implications (Kwan & Hong, 1998; Arentze & Timmermans, 2005).

Deterministic constraint-based approaches have been used in the literature to model choice sets for example in the context of mode choice (Ben-Akiva & Lerman, 1974), recreational site choice (Termansen, McClean, & Skov-Petersen, 2004), destination choice (Scott, 2006), and residential location choice (Zolfaghari, 2013). These methods assume that households/individuals use non-compensatory decision rules for screening of alternatives based on some behavioural constraints. Alternatives are removed from

the individual choice set when certain attributes of an alternative exceed exogenous thresholds. These exogenous thresholds can be either imposed deterministically based on insights from the data (Farooq & Miller, 2012) or can be computed (Zolfaghari, 2013). Importance sampling techniques have also been used in the context of residential location choice modelling (e.g., Rashidi et. Al, 2012; Zolfaghari, 2013, etc.). These techniques are similar to the deterministic constraint-based approaches but allow proportional sampling of alternatives from within and outside the threshold zone. For example, Farooq and Miller (2012) applied importance sampling to construct individual choice sets for residential location choice modelling by taking 75% of alternatives within 15 km of the past location and the remaining 25% from outside the threshold. Since these techniques are based on assumptions made by the analyst, there is a high risk of choice set misspecification and consequently, poor model fit and biased parameter estimation.

Heuristic-based semi-compensatory approaches can avoid the combinatorial number of choice sets in the probabilistic approach and the risk of elimination by aspect approaches (e.g., ignoring alternatives that have non-zero choice probabilities), and therefore become appealing in case of modelling with a large universal choice set. The basic principle of these methods is the adjustment of systematic utility based on the probability of an alternative being in the individual choice set. A penalty term is introduced in the utility equation for the adjustment. Therefore, the utility function for alternative i and person n can be defined as follows:

$$U_{in} = V_{in} + ln(\phi_{in}) + \varepsilon_{in} \qquad \text{where } 0 \leq \phi in \leq 1 \tag{2}$$

where ϕ_{in} is the probability of alternative i being in the choice set of individual n, V_{in} is the deterministic utility of the alternative i for individual n and ε_{in} is the usual identically and independently distributed (iid) extreme value error term. If there is a full probability of an alternative being in the individual choice set, the penalty term becomes zero (i.e., no adjustment is required).

Different functional forms of ϕ_{in} have been used in different semi-compensatory approaches. ϕ_{in} is expressed as a binary logit function of attributes related to choice set formation in the Implicit Availability Perception Random Utility (IAPRU) model proposed by Cascetta and Papola (2001). The mathematical expression of ϕ_{in} is as follows:

$$\phi_{in} = \frac{1}{1 + \exp \sum_k (-\mu_k z_{ik})} \tag{3}$$

where z_{ik} is a parameter associated with attribute k and alternative i and μ_k is the scale parameter. Second order utility penalization is also proposed based on the Taylor series expansion for further utility cut-off of less attractive alternatives (Cascetta & Papola, 2001). This method, however, leads to estimation difficulties in complex specifications with multiple constraints. Moreover, second order utility penalization has convergence issues if a chosen alternative is subjected to an extreme penalty. For example, if commute distance is considered as an availability/perception attribute (k) in residential location choice modelling and the penalty parameter (μ_k) is 0.8, the utility cut-off for an alternative 10 km away from the individual workplace is 1498 units, leading to a choice probability close to zero. If that alternative is chosen by anyone (which is possible in reality if the alternative is attractive in terms of all other attributes), it can lead to estimation problems.

In a simpler method, Cascetta and Papola (2009) proposed to simulate the choice set (i.e., availability) implicitly based on the rule of dominance among alternatives. The principle is that an alternative i dominates alternative j if j is worse than i with respect to dominant attributes K (where K can be a single

or multiple attributes). An alternative *j* is worse than *i* if quality attributes Q_i (i.e., attributes with positive coefficients) are smaller in *j* than in *i* and cost attributes C_i (i.e., attributes with negative coefficients) are larger in *j* than in *i*, with at least one inequality strictly satisfied. This framework, however, does not account for how worse one alternative is compared to another alternative. The rule of dominance can be expressed as follows:

$$y_{ij}^{n} = \begin{cases} 1 \;\; if \; Q_{ink} > Q_{jnk} \; and \; C_{ink} < C_{jnk}, \; \forall k \in K \\ 0 \quad\quad otherwise \end{cases} \tag{4}$$

y_{ij}^{n} indicates that alternative *j* is dominated by alternative *i* for individual *n* if inequality criteria is satisfied for at least one attribute. The penalty term can be expressed as follows:

$$\ln(\phi_{in}) = \mu\left(\sum_{i \in C} y_{ij}^{n}\right), \quad \mu \; is \; the \; scale \; parameter, \; \mu < 0 \tag{5}$$

The constrained multinomial logit model (CMNL) proposed by Martínez et al. (2009) has greater flexibility to accommodate multiple constraints with exogenous bounds (both upper and lower) to simulate the individual choice set implicitly. A binary logit functional form is also considered here to estimate the probability of alternatives being in the individual choice set (ϕ_{in}). The binary logit functional form of ϕ_{in} with an upper threshold t_{nk} on a constrained attribute *k* can be presented as follows:

$$\phi_{ink} = \frac{1}{1 + \exp\left(\mu_k(z_{ik} - t_{nk} + \delta_k)\right)} \tag{6}$$

$$\delta_k = \frac{1}{\mu_k} \ln\left(\frac{1 - \eta_k}{\eta_k}\right) \tag{7}$$

where z_{ik} is the value of constrained attribute *k*, t_{nk} is the threshold of attributes *k* for individual *n*, η_k is the cut off tolerance (proportion of decision makers violating the threshold) and μ_k is the scale parameters ($\mu_k > 0$). ϕ_{ink} collapses to 1 and η_k under the conditions $(z_{ik} - t_{nk}) = -\infty$ and $z_{ik} = t_{nk}$, respectively. If a constraint is applied on multiple attributes, the total penalty term becomes

$$\ln(\phi_{in}) = ln\left[\prod_{k=1}^{K} \phi_{ink}\right] = \sum_{k=1}^{K} ln\left(\phi_{ink}\right) = -\sum_{k=1}^{K} \ln(1 + \exp(\mu_k(z_{ik} - t_{nk} + \delta_k))) \tag{8}$$

The exogenous threshold-based heuristic adopted in the CMNL model is relevant for many cases. For instance, it is unlikely that a low-income household considers very expensive houses as options. Therefore, the CMNL model has received considerable attention recently and found wider application in literature for example in the context of modelling location choice (Martínez & Hurtubia, 2006), parking management (Caicedo, Lopez-Ospina & Pablo-Malagrida, 2016), mode choice (Castro, Martínez, & Munizaga 2013). However, Bierlaire, Hurtubia and Flötteröd, (2010) demonstrated the inconsistency of the choice set generated in the CMNL model with the Manski framework using simulated experiments on synthetic data.

Paleti (2015) proposed the r[th]-order CMNL model (called rCMNL) where the complexity is linear with the size of the choice set. A higher order functional form of the CMNL penalty term (ϕ_{in}) is proposed in this regard. The r[th]-order penalty in rCMNL is the natural logarithm of the following r[th]- order expressions.

$$\phi_{in}^{1} = \phi_{in} \;\; First\ order \tag{9}$$

$$\phi_{in}^{2} = \phi_{in}[(1 - \bar{P}_{in}) + \phi_{in}^{1} \times \bar{P}_{in}] \text{ Second order} \tag{10}$$

$$\phi_{in}^{3} = \phi_{in}[(1 - \bar{P}_{in}) + \phi_{in}^{2} \times \bar{P}_{in}] \;\; \text{Third order} \tag{11}$$

$$\phi_{in}^{r} = \phi_{in}[(1 - \bar{P}_{in}) + \phi_{in}^{r-1} \times \bar{P}_{in}] \text{ where } \phi_{in}^{0} = 1 \quad r^{th}\text{-}order \tag{12}$$

Where $\bar{P}_{in}$ is the probability of choosing alternative *i* from the full choice set without any penalization.

$$\bar{P}_{in} = \frac{e^{V_{in}}}{\sum_{j \in C} e^{V_{jn}}} \tag{13}$$

Using synthetic data and real-world data, the author demonstrates that higher order penalization performs considerably better than the CMNL model in terms of replicating the Manski model parameters. However, in both examples, the number of alternatives in the universal choice sets was very limited (three and five alternatives, respectively). If the number of alternatives in the choice increases, the probability of each alternative is likely to go down. For a very large universal choice set (hundreds to thousands of alternatives), $\bar{P}_{in}$ will be too small and $\phi_{in}^{r} \simeq \phi_{in}$ (i.e., the model collapses to the first order CMNL).

3 Improved constrained multinomial logit model (ICMNL)

Though the complexity of the CMNL model remains linear with an increase in the number of alternatives in the choice set, it struggles to replicate the outcomes of the Manski method (Bierlaire, Hurtubia, & Flötteröd, 2010). The penalty term considered in the CMNL model is a first order penalty derived from the attributes that influence individual choice sets. The higher order utility penalization proposed in the rCMNL model can minimize the error in the CMNL model outcomes when the size of the universal choice set is small. In case of a large universal choice set, the higher order penalty in the rCMNL model collapses to the first order CMNL penalty and cannot offer any further improvement. This is due to the fact that the rCMNL penalty depends on the probability of choosing alternatives from the universal choice set which is likely to be very small for large universal choice sets. We therefore propose an alternate formulation of higher order approximation of the availability term (ϕ_{in}) in the CMNL model based of the concept of a Taylor series expansion which is independent of the number of alternatives in the universal choice set. This is motivated by the application of Taylor series expansion in the context of the Implicit Availability Perception (IAP) logit model (see Cascetta & Papola, 2001, for details). The basic utility equation with implicit availability of alternatives can be expressed as follows:

$$U_{in} = V_{in} + \ln(\phi_{in}) + \varepsilon_{in} \quad \text{where } 0 \leq \phi_{in} \leq 1 \tag{14}$$

In the CMNL model, ϕ_{in} is estimated as a binary logit function of constrained attributes. Since the constrained attributes and exogenous thresholds considered in the CMNL model are based on the re-

searcher's assumptions, the estimated value of the implicit availability term (ϕ_{in}) is unlikely to be the true value. Therefore, in the ICMNL model, we have decomposed the availability into expected availability and an error term. Thus, the true penalty can be expressed as $\ln(\phi_{in})=E(\ln(\phi_{in}))+\delta_{in}$ where δ_{in} is the error term (divergence between true and expected penalty). The utility equation can be revised as follows

$$\begin{aligned} U_{in} &= V_{in} + E(\ln(\phi_{in})) + \delta_{in} + \varepsilon_{in} \\ U_{in} &= V_{in} + E(\ln(\phi_{in})) + \tau_{in} \end{aligned} \tag{15}$$

For simplicity, the total error (τ_{in}) is assumed to be independently and identically distributed (IID). Based on the 2nd order Taylor series expansion, the expected penalty can be expressed as below

$$\begin{aligned} E(\ln(\phi_{in})) &= E(\ln\bar{\phi}_{in}) + E\left(\frac{\phi_{in}-\bar{\phi}_{in}}{\bar{\phi}_{in}}\right) - E\left(\frac{(\phi_{in}-\bar{\phi}_{in})^2}{2(\bar{\phi}_{in})^2}\right) \\ &= \ln(\bar{\phi}_{in}) - \frac{Var(\phi_{in})}{2(\bar{\phi}_{in})^2} \end{aligned} \tag{16}$$

Since the distribution of ϕ_{in} is unknown, the variance of ϕ_{in} is also unknown. Considering the variance of the Bernoulli distribution, Var (ϕ_{in})=$\bar{\phi}_{in}$ (1-$\bar{\phi}_{in}$) , the equation (16) can be modified as follows

$$E(\ln(\phi_{in})) = \ln\bar{\phi}_{in} - \frac{(1-\bar{\phi}_{in})}{2\bar{\phi}_{in}} \tag{17}$$

The utility equation (15) can be presented as follows

$$\begin{aligned} U_{in} &= V_{in} + \ln\bar{\phi}_{in} - \frac{(1-\bar{\phi}_{in})}{2\bar{\phi}_{in}} + \tau_{in} \\ &= V_{in} + \ln\bar{\phi}_{in}^{T^{(2)}} + \tau_{in} \end{aligned} \tag{18}$$

$\ln\bar{\phi}_{in}^{T^{(2)}}$ is the second order utility penalty where

$$\bar{\phi}_{in}^{T^{(2)}} = \bar{\phi}_{in} * e^{-}\left(\frac{1-\bar{\phi}_{in}}{2\bar{\phi}_{in}}\right), \tag{19}$$

The average availability $\bar{\phi}_{in}$ can be estimated implicitly using the mathematical formulation proposed in the CMNL model (see equation 6). A constraint is applied to the attributes related to the alternative to estimate the choice set probability of the alternative. If the constraint is applied on multiple attributes, the total penalty becomes:

$$\begin{aligned} \ln(\bar{\phi}_{in}^{T^{(2)}}) &= ln\left[\prod_{k=1}^{K}\bar{\phi}_{ink}^{T^{(2)}}\right] = \sum_{k=1}^{K} ln(\bar{\phi}_{ink}^{T^{(2)}}) \\ &= -\sum_{k=1}^{K}\left\{\ln\left(1+\exp(\mu_k(z_{ik}-t_{nk}+\delta_k))\right) + \frac{\exp(\mu_k(z_{ik}-t_{nk}+\delta_k))}{2}\right\}\Big\} \end{aligned} \tag{20}$$

Where $\bar{\phi}_{ink}^{T^{(2)}}$ represents the probability of alternative *i* to be in the choice set of individual *n* when a constraint is applied on attribute *k*. Therefore, the probability of choosing alternative *i* by household *n* in the ICMNL model is

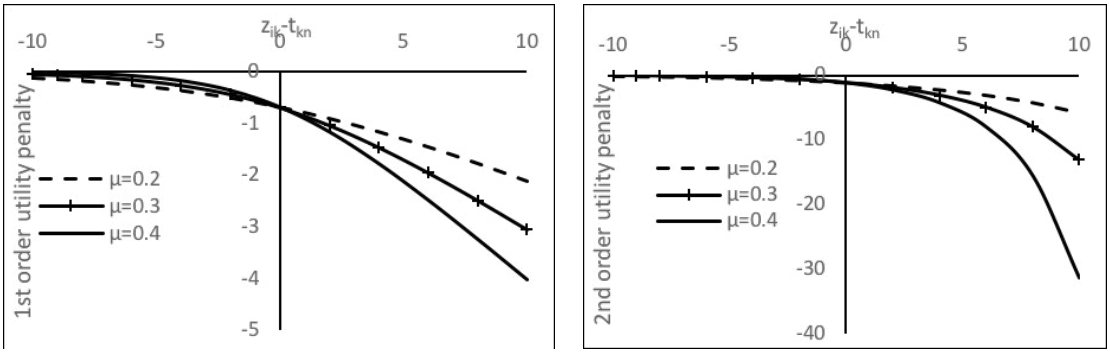
$$P_{in} = \frac{e^{V_{in}+\ln(\bar{\phi}_{in}^{T(2)})}}{\sum_{j\in C} e^{V_{jn}+\ln(\bar{\phi}_{jn}^{T(2)})}} \tag{21}$$

and the log likelihood function is

$$LL(\beta) = \sum_n \sum_i y_{in} \ln(P_{in}) \tag{22}$$

where y_{in} = 1 if alternative i is chosen by household n and y_{in} = 0 for all nonchosen alternatives. The maximum likelihood estimates of the model parameters are found by maximizing this function. Functional forms used in different methods are explained in the following sections.

If the attributes move away from the bound, the rate of increment of 2nd order penalties in ICMNL become considerably stronger than the 1st order penalty. For example, for μ=0.4, if the value of $(z_{ik} - t_{nk})$ moves from 5 to 10, the increment of the first-order penalty is 2 units which is 25 units for second order penalty (Figure 1). Therefore, the first order penalty is considered as a soft penalty and the second order penalty is considered as a hard penalty. The scale parameter also determines the size of the penalty.



(a) 1st order penalty (soft) (b) 2nd order penalty (hard)

Figure 1. Penalization of the utility function

Due to applying hard penalties on those alternatives that are unlikely to be in the individual choice set, the choice probabilities of these alternatives tend to zero which is behaviourally reasonable. Therefore, the ICMNL model is expected to be a better approximation of the Manski formulation. The performance of the ICMNL is first evaluated here with a simple analysis (similar to that in Bierlaire, Hurtubia, & Flötteröd, 2010). For this analysis, only two alternatives are considered where alternative 1 is always available in the choice set (ϕ_1=1) and alternative 2 has a probability of being in the choice set ($\phi_2 \leq 1$). This hypothesis is similar to the CMNL concept where alternatives within a threshold are always available in the choice set and choice set membership probabilities are assigned for those alternatives that are outside the threshold zone. In the CMNL, rCMNL, and ICMNL, the probability of choosing alternative 1 is as follows:

$$P_1 = \frac{e^{V_1+\ln(\phi_1')}}{e^{V_1+\ln(\phi_1')} + e^{V_1+\ln(\phi_2')}} \tag{23}$$

$$P_1 = \frac{e^{V_1}}{e^{V_1}+e^{V_2+\ln(\phi_2')}}, \; since \; \phi_1 = 1 \tag{24}$$

where V_1 and V_2 are the systematic utilities of alternatives 1 and 2, respectively.

The mathematical formulation of penalty terms in the CMNL, rCMNL and ICMNL models can thus be summarized as follows:

$$\phi_2' = \phi_2 \qquad \text{CMNL} \tag{25}$$

$$\phi_2' = \phi_2[(1 - \bar{P}_2) + \phi_2 \times \bar{P}_2] \qquad 2^{nd} \text{ order of rCMNL} \tag{26}$$

$$\phi_2' = \phi_2 * e^{-\left(\frac{1-\phi_2}{\phi_2}\right)} \qquad \text{ICMNL} \tag{27}$$

The probability of choosing alternative 1 based on the Manski formulation is as follows:

$$P_1 = P(C[1]) * \frac{e^{V_1}}{e^{V_1}} + P(C[1,2]) * \frac{e^{V_1}}{e^{V_1}+e^{V_2}} \tag{28}$$

where P(C[1]) and P(C[1,2]) are the probabilities of choice sets containing alternative 1 only and both of the alternatives (1 and 2), respectively. The probability of a given choice set can be expressed as follows (Bierlaire, Hurtubia, & Flötteröd 2010).

$$P(C[1]) = \frac{\phi_1(1-\phi_2)}{1-(1-\phi_1)(1-\phi_2)} = 1 - \phi_2, since\ \phi_1 = 1 \tag{29}$$

$$P(C[1,2]) = \frac{\phi_1\phi_2}{1-(1-\phi_1)(1-\phi_2)} = \phi_2, since\ \phi_1 = 1 \tag{30}$$

$$\text{Therefore } P_1 = (1 - \phi_2) + \phi_2 * \frac{e^{V_1}}{e^{V_1}+e^{V_2}} \tag{31}$$

The choice probability of alternative 1 (P1) is calculated for different values of ϕ_2 (probability of alternative 2 being in the choice set) under different conditions using the CMNL, rCMNL, ICMNL and Manski formulation presented above. The results are plotted in Figure 2.

- Under condition (a) when the utility of alternative 1 (V_1) is larger than the utility of alternative 2, the semi-compensatory approaches (CMNL, rCMNL and ICMNL model) can replicate the Manski model probability quite well though the CMNL model gives the best fit (Figure 2a).
- Under condition (b) when the utility of both of the alternatives is the same or close to each other, CMNL, rCMNL and ICMNL model can still offer a close approximation to the Manski method though rCMNL model results in the best fit (Figure 2b).
- The main concerns are conditions (c) and (d), when the utility of alternative 1 is smaller than alternative 2. In these cases, the semi-compensatory approaches cannot reproduce the Manski model results and the error produced by the semi-compensatory approaches becomes larger with the decrease of the utility of alternative 1 (also demonstrated by Bierlaire, Hurtubia, & Flötteröd, 2010). However, the ICMNL model can considerably reduce the error between the Manski and the CMNL model (Figures 2c and 2d).[1,2]

[1] ICMNL model has asymmetric domains of under-estimating and over-estimating the Manski model probabilities. The asymmetric property of under and overestimation depends on the differences in the utility of the alternatives. Since the ICMNL model can reduce the error considerably compared to the CMNL and the rCMNL models (Figure 2c and 2d), this asymmetric property is unlikely to affect the model performance.

[2] Under condition (d), the CMNL and the rCMNL models produce same results, therefore, their plots in Figure 2d overlap.

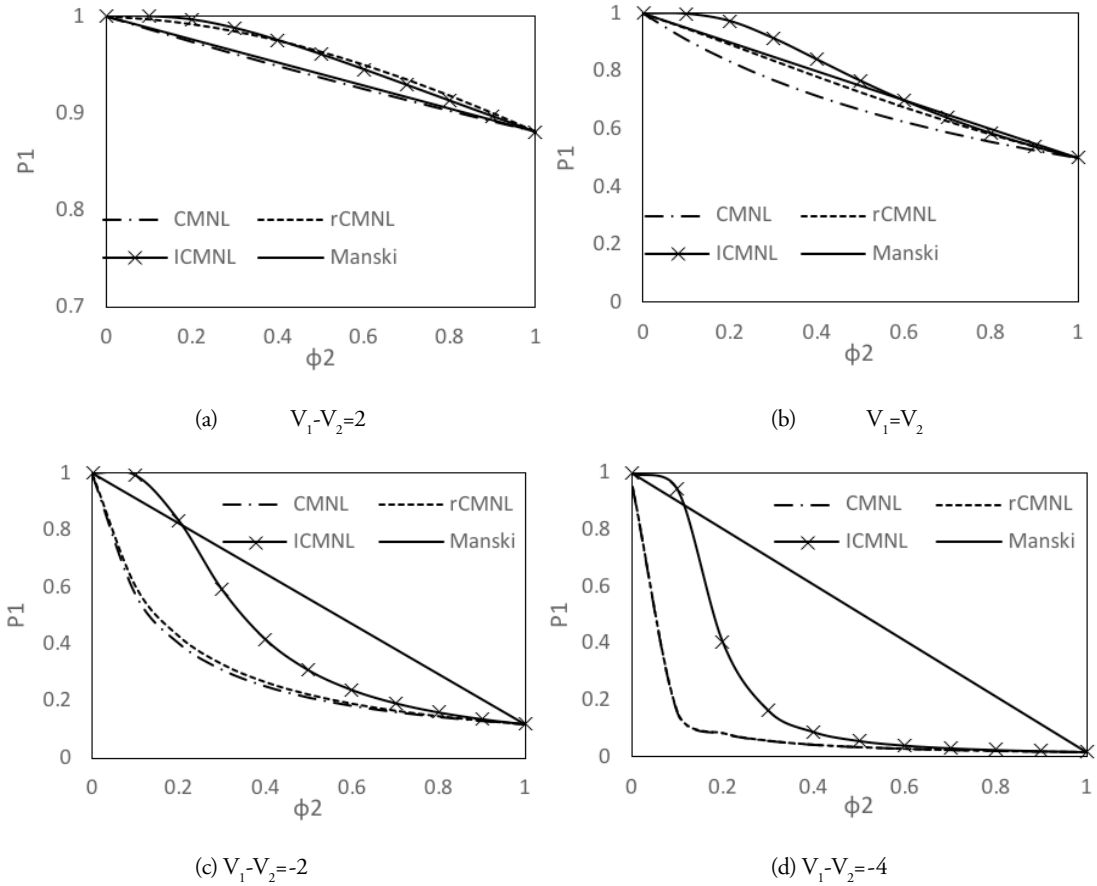


Figure 2. Choice probability of alternative 1 for different utility differences

4 Data analysis and variable specification

The London Household Survey Data (LHSD) collected in 2002 is considered as the primary data source for this study. This dataset contains detailed information about the socio-demographic characteristics (household size, income, etc), dwelling characteristics (tenure type, size, price/rent, etc.), employment status, home and work location, car ownership, etc of 8,158 households from the Greater London Area (GLA). The dataset contains information of 4,491 households living in occupier-owned houses, 2,489 households living in houses rented from councils or housing authorities, 1,087 households living in privately rented houses and 91 households living in shared accommodation. Since this study only considered households that live in owned or privately rented houses, have at least one working member and had at least one residential move within the GLA, the final dataset contains information from 1,875 owners and 382 renters.

The Ward Atlas Data (WAD) for 2002 is also considered for zone level aggregated demographic, land use and other information. The origin-destination (OD) matrix of GLA obtained from the London Transport Studies (LTS) model is used to extract the distances of the new alternate locations from the past home location, work location and the central business district (CBD). Combining all these data sets poses significant challenges due to the consideration of different geographical boundaries in the different datasets. GIS-based conversion to get unique geographical boundaries across different data sets for combining them helps to minimize the error. Merging several datasets also allows us to test a large set of parameters in the models. Parameters considered for this study are listed in Table 1. Details about the data processing and data characteristics are available in Haque, Choudhury and Hess (2018).

Table 1. Parameters considered for this study

Name of the explanatory variables	Interaction variables	Data Sources	Unit	Anticipated impact
Dwelling characteristics				
Dwelling cost	Low income	LHSD &WAD	Pound	-
	Middle income	LHSD &WAD	Pound	-
	High income	LHSD &WAD	Pound	-
Dwelling type				
Detached house	Inner London	LHSD &WAD	Percentage	-
	Outer London	LHSD &WAD	Percentage	-
Flat house	Inner London	LHSD &WAD	Percentage	+
	Outer London	LHSD &WAD	Percentage	-
Location and land use characteristics				
Land use type				
Residential land use	Inner London	LHSD &WAD	Percentage	+
	Outer London	LHSD &WAD	Percentage	+
Commercial land use	-	LHSD &WAD	Percentage	-
Land use mix	-	LHSD &WAD	Index varies from 0 to 1	+
Ethnicity	White	LHSD &WAD	Percentage	+
	Asian	LHSD &WAD	Percentage	+
	Black	LHSD &WAD	Percentage	+
Dwelling density	Inner London	LHSD &WAD	Per square KM	-
	Outer London	LHSD &WAD	Per square KM	-
School quality	School going child	LHSD &WAD	Unitless score	+
Crime rate	-	LHSD &WAD	Per thousand population	-
Household size	-	LHSD &WAD	Number	+/-
Employment opportunity	-	LHSD &WAD	Per person	+
Distance from CBD	-	LHSD &LTS	Kilometre	+
Distance from past home	-	LHSD &LTS	Kilometre	-
Transport and travel characteristics				
Public transport accessibility	Having cars	LHSD &WAD	Score out of 8	+/-
	Don't have car	LHSD &WAD	Score out of 8	+
Commute distance	-	LHSD &LTS	Kilometre	-
Constants	Central London	-	-	+/-
	North London	-	-	+/-
	South London	-	-	+/-
	East London	-	-	+/-
	West London	-	-	+/-

Not surprisingly, statistical analysis of the data shows that households are inclined to choose residential location alternatives close to their current home. However, owners' preferences to relocate near to their current home are found to be stronger than those of renters (Figure 3). For example, 90% owners chose their new locations within 14 km of their past homes and for the remaining 10%, they are spread between 14 and 50km whereas 90% renters chose their new locations within 18 km of their past homes

and the remaining 10% are spread between 18 and 50km. Sharp slope changes of the curves at a certain point in Figure 3 indicate the possible threshold effects of choice set consideration. Since most of the households chose their new locations close to their past home (e.g., 90% owners chose within 14 km of past home), it is unlikely that they considered the alternatives far from their past home (outside a threshold zone).

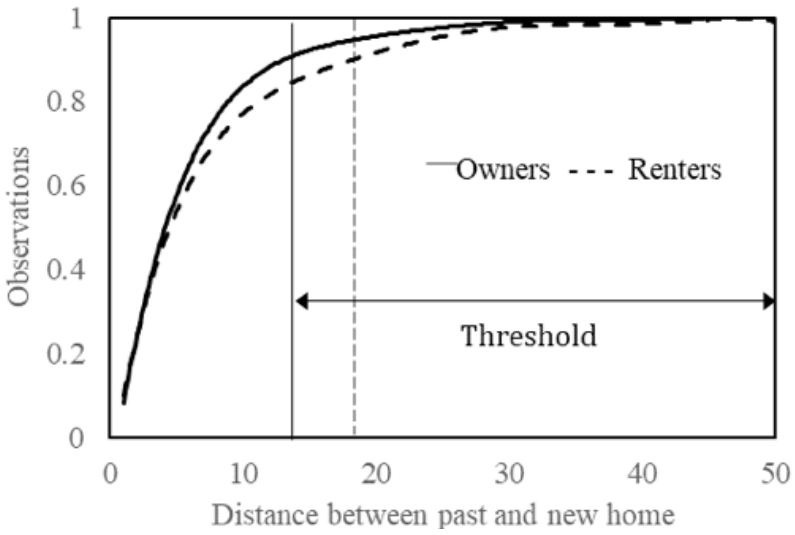


Figure 3. Distance between past and new home

5 Estimation results

Residential location choices of owners and renters are modelled in this study using the existing (CMNL and rCMNL models) and proposed (ICMNL model) semi-compensatory approaches where choice sets are simulated implicitly based on exogenous constraints on attributes. Although the choice set of an individual is likely to be influenced by a set of parameters (e.g., commute distance, distance of alternatives from the past home, distance of alternatives from the CBD, housing cost, etc.), households are found to have strong preferences to the alternatives close to their past home locations and work locations based on statistical analysis of the data. Therefore, we have explicitly tested the influence of distance of alternatives from the past home location and commute distance on individual choice sets. Only distance from the past home is found to have a significant influence on household choice set consideration. An exogenous threshold is applied to the parameter of past home distance to simulate the choice set. Different thresholds are tested in the models and the threshold value resulting in the maximum likelihood is considered for the final model.

As an explanatory variable of individual choice, a large set of parameters including location characteristics, aggregate level dwelling chrematistics, commute characteristics and interaction variables are considered in the models. Potential correlations across the model parameters (independent variables) are tested and very weak associations are observed across them. For instance, the correlation value between commute distance and distance from the CBD is found to be 0.19 and 0.09, for owners and renters, respectively. Different model specifications are tested and the final models contain the parameters that

are statistically significant in at least one of the models.[3] Several higher order approximations of the rCMNL model were tested in this study and the 3rd order approximation was found to give stable results in terms of improvement in model fit. The performance of the models estimated using the CMNL, rCMNL and ICMNL techniques was analysed based on the improvement of log-likelihood in the estimation sample.[4]

5.1 Ownership

The estimated parameters of the ownership models are presented in Table 2. It is observed that the ICMNL model shows significant improvement in log likelihood over the CMNL (146 units) and the rCMNL models (143.6 units). However, the improvement of the rCMNL model over the CMNL model is insignificant (only 2.4 units). This is due to the fact we alluded to in the earlier section in that the rCMNL model is equivalent to the CMNL model when the size of the universal choice set is large. Estimated parameters are found to be stable across the models estimated using different techniques.

All the parameters considered in the models have the expected signs and most of them are found to be statistically significant. Household cost sensitivity is found to be heterogeneous across different income groups. For example, the lower income group is more price sensitive than the higher income group, as expected. Preferences for ethnic similarity (where a higher number of households come from the same ethnic group) are found to have a positive and statistically significant effect. Results also show that households dislike higher levels of dwelling density, commercial activities and crime in their residential areas. Although households prefer to live in areas with higher residential activities, they also prefer areas with more balanced land use patterns. Households do not prefer an area with a higher percentage of detached houses, this may be due to the excess price of detached houses in GLA, even after accounting for price in the model. However, households are found to be inclined to buy flats in inner London areas and seem to dislike buying flats in outer London areas, all else being equal. Households are also found to prefer areas having greater employment opportunities, good school facilities and those further from the central business district (CBD). The household size (absolute difference between individual household size and zonal average) parameter shows a negative effect on utility. Increases in public transport accessibility increase the utility of 'car-less' households but decrease the utility of 'car-owning' households. It is also observed that increased commute distance adds disutility to the residential location alternatives.

5.2 Renting

The goodness of fit of the proposed ICMNL model is found to be better than that of the CMNL model and the rCMNL model also in the renting dataset (Table 3). However, a loss of likelihood has been observed in the rCMNL model here compared to the CMNL model. In the rCMNL model, a larger penalty is applied to the alternatives having higher choice probabilities without the penalty term. Therefore, chosen alternatives outside the threshold zone are likely to be assigned a large penalty resulting in a decrease in the model fit.

---

[3] For investigating the differences in the choice set consideration of both owners and renters, the same set of parameters are considered in the ownership and renting models although few of them are insignificant in one model but significant in another model. To check whether the insignificant parameters are affecting the results, we have estimated the models ignoring the insignificant parameters and found no significant differences in the estimation results.

[4] Goodness of fit of the heuristic based semi compensatory approaches (e.g., IAPRU model, dominance rule-based approach, CMNL, etc.) can vary case by case. It depends on the appropriateness of the heuristic for the specific context. For example, dominance rule based approaches might be suitable for one case and exogenous threshold based approaches (CMNL) could perform better in another context. It is difficult to compare the methods based on different heuristics in a single dataset and to draw a general conclusion. Therefore, we only compared the performances of CMNL, rCMNL and ICMNL model in this study where this problem does not arise.

Table 2. Estimation results of residential ownership models

Parameters	CMNL		rCMNL		ICMNL	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
Constants						
Central London	0.3303	2.5	0.3369	2.5	0.4193	3.0
South London	0.3585	3.4	0.3625	3.4	0.3849	3.3
North London	0.6142	4.7	0.6229	4.7	0.6644	4.7
East London	0.8201	6.6	0.8307	6.6	0.9424	6.9
Dwelling characteristics						
Dwelling cost (price* 0.0001)						
Income less than £30,000	-0.5612	-6.9	-0.5650	-7.0	-0.6092	-7.2
Income between £30,000 to £60,000	-0.4624	-6.4	-0.4655	-6.4	-0.5049	-6.7
Income more than £60,000	-0.2373	-4.7	-0.2390	-4.7	-0.255	-4.8
Missing values	-0.0473	-1.4	-0.0475	-1.4	-0.0611	-1.8
Dwelling type						
Detached house in inner London	-0.1297	-4.8	-0.1305	-4.9	-0.1235	-4.6
Detached house in outer London	-0.0294	-5.8	-0.0297	-5.8	-0.0302	-5.9
Flat in inner London	0.0226	5.2	0.0227	5.2	0.0218	5.0
Flat in outer London	-0.0076	-2.7	-0.0077	-2.7	-0.0067	-2.4
Location and land use characteristics						
Land use type						
Residential land area in inner London	0.1169	7.7	0.1177	7.7	0.1077	7.0
Residential land area in outer London	0.1833	8.3	0.1847	8.3	0.1826	8.2
Commercial land area in inner and outer London	-0.0478	-4.4	-0.0480	-4.4	-0.0430	-3.9
Land use mix	1.0401	3.0	1.0476	3.0	0.8134	2.4
Ethnic composition						
Ratio of white people × white dummy	0.0183	8.1	0.0185	8.2	0.0186	8.1
Ratio of asian people × asian dummy	0.0338	8.7	0.0344	8.8	0.0323	8.0
Ratio of Black people × black dummy	0.0451	5.7	0.0456	5.8	0.0428	5.3
Dwelling density						
Inner London	-0.0116	-2.3	-0.0117	-2.3	-0.0110	-2.1
Outer London	-0.0993	-9.7	-0.1000	-9.8	-0.1001	-9.8
School quality	0.0032	2.1	0.0033	2.1	0.0025	1.6
Crime rate	-0.0333	-0.6	-0.0356	-0.7	-0.0328	-0.6
Household size	-0.2816	-2.9	-0.2854	-3.0	-0.2722	-2.8
Employment opportunity	0.1147	2.4	0.1179	2.4	0.1219	2.5
Distance from CBD	0.0871	10.0	0.0883	10.1	0.1026	11.2
Transport and travel characteristics						
Public transport accessibility						
Households own car	-0.1963	-4.0	-0.1980	-4.0	-0.2094	-4.2
Households do not own car	0.1133	1.5	0.1152	1.6	0.0920	1.2
Commute distance	-0.1478	-28.8	-0.1493	-29.0	-0.1537	-26.7
Penalty parameter (μ)						
Distance from past home	0.2001	38.2	0.2023	38.6	0.0190	34.7
Measures of model fit						
Number of observations	1875		1875		1875	
Initial LL	-11644.8751		-11644.8751		-11644.8751	
Final LL	-7744.6590		-7742.2800		-7598.6130	
Adjusted ρ2	0.332		0.333		0.345	

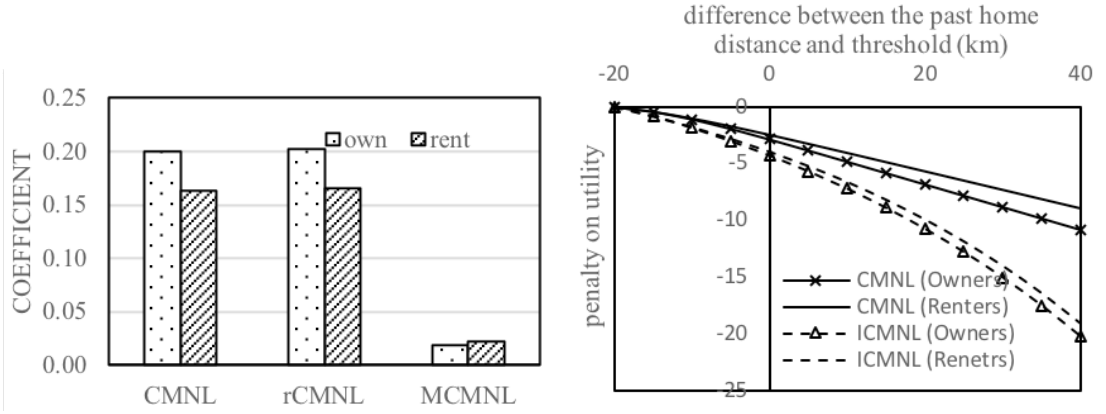
Table 3. Estimation results of residential renting models

Parameters	CMNL		rCMNL		ICMNL	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
Constants						
Central London	0.1510	0.6	0.1553	0.6	0.1860	0.8
South London	0.3258	1.4	0.3293	1.4	0.3292	1.4
North London	0.5687	2.0	0.5764	2.1	0.6112	2.2
East London	0.6640	2.5	0.6731	2.7	0.7213	2.7
Dwelling characteristics						
Dwelling cost (monthly rent* 0.01)						
Income less than £30,000	-0.2634	-5.2	-0.2658	-5.2	-0.2721	-5.3
Income between £30,000 to £60,000	-0.1204	-2.1	-0.1216	-2.2	-0.1172	-2.1
Income more than £60,000	-0.0963	-2.7	-0.0972	-2.8	-0.1010	-2.8
Missing values	-0.0793	-2.8	-0.0801	-2.8	-0.0830	-2.8
Dwelling type						
Detached house in inner London	-0.0250	-0.6	-0.0250	-0.6	-0.0273	-0.7
Detached house in outer London	-0.0143	-1.0	-0.0145	-1.0	-0.0144	-1.0
Flat in inner London	0.0290	3.5	0.0292	3.5	0.0295	3.5
Flat in outer London	-0.0011	-0.2	-0.0010	-0.2	-0.0009	-0.1
Location and land use characteristics						
Land use type						
Residential land area in inner London	0.1327	5.0	0.1340	5.1	0.1294	4.9
Residential land area in outer London	0.2018	4.0	0.2037	4.1	0.1966	3.9
Commercial land area in inner and outer London	-0.0607	-3.1	-0.0614	-3.2	-0.0590	-3.1
Land use mix	2.0014	2.1	2.0274	2.4	1.8599	2.0
Ethnic composition						
Ratio of white people × white dummy	0.0199	3.7	0.0201	3.7	0.0200	3.7
Ratio of asian people × asian dummy	0.0439	5.2	0.0447	5.3	0.0451	5.2
Ratio of Black people × black dummy	0.0382	2.6	0.0386	2.6	0.0370	2.5
Dwelling density						
Inner London	-0.0083	-1.1	-0.0083	-1.1	-0.0085	-1.1
Outer London	-0.1013	-4.6	-0.1022	-4.6	-0.1000	-4.5
School quality	0.0022	0.5	0.0023	0.5	0.0016	0.3
Crime rate	-0.2051	-2.2	-0.2086	-2.3	-0.2077	-2.3
Household size	-0.0957	-0.5	-0.0966	-0.5	-0.0947	-0.5
Employment opportunity	0.3094	3.6	0.3142	3.7	0.3147	3.7
Distance from CBD	0.0557	2.8	0.0566	2.8	0.0660	3.2
Transport and travel characteristics						
Public transport accessibility						
Households own car	0.0416	0.4	0.0421	0.4	0.0351	0.3
Households do not own car	0.2572	2.5	0.2592	2.5	0.2479	2.4
Commute distance	-0.1872	-16.4	-0.1890	-16.6	-0.1961	-16.2
Penalty parameter (μ)						
Distance from past home	0.1631	14.4	0.1651	14.5	0.0227	14.7
Measures of model fit						
Number of observations	382		382		382	
Initial LL	-2372.4492		-2372.4492		-2372.4492	
Final LL	-1678.7620		-1679.4270		-1662.8590	
Adjusted ρ2	0.322		0.322		0.325	

All parameters in the renting models also obtain the expected sign. Some of the estimated parameters are found to be statistically insignificant but are retained in the models to ensure consistent parameter specification in both ownership and renting models. Since the estimated parameters in the renting models give the same sign as the corresponding parameters estimated in ownership models, the interpretations are the same.

5.3 Contrast between ownership and renting

In terms of the penalty term in the models to simulate the choice set probability implicitly, our results show preferences consistent with the earlier statistical analysis (owners' preference for alternatives close to the past home location is stronger than renters' preference, Figure 1). The differences in the estimated values of the penalty term (μ) in ownership and renting models are found to be statistically significant (Figure 4a). Figure 4b also confirms that the penalty applied on owners' utility due to the increase of distance of alternatives from their past home is always higher than that of renters. This means that alternatives close to the current home have a higher probability to be included in the choice set of owners than renters. The direction of sensitivity (sign) of the explanatory parameters in the compensatory utility is found to be consistent both in the ownership and renting models but the sensitivity of several parameters are found to be significantly different in both models (e.g., commute distance, distance from CBD, etc.).



(a) Estimated values of penalty parameters

(b) Utility penalization

Figure 4. Impact of penalty terms on owners' and renters' choices

6 Validation results

For model validation, both the ownership and renting datasets are randomly divided into five rolling subsets. Each subset consists of 80% of the data for estimation and 20% for validation. Models are estimated for five estimation subsets of owners and five estimation subsets of renters. It is observed that the estimated parameters are consistent across the models estimated using the different subsets of the owner and renter data. In terms of the goodness of fit of the estimated models, the ICMNL model consistently shows the best performance in all subsets, both for ownership and renting (Tables 4 and 5 respectively).

Table 4. Final log-likelihood of models estimated for estimation subsets of owners data

Subset	Number of Observations	Initial LL	Final LL		
			CMNL	rCMNL	ICMNL
Subset1	1500	-9315.90	-6173.75	-6173.42	-6048.39
Subset2	1500	-9315.90	-6207.40	-6207.11	-6091.69
Subset3	1500	-9315.90	-6195.84	-6195.24	-6079.12
Subset4	1500	-9315.90	-6224.88	-6224.35	-6107.48
Subset5	1500	-9315.90	-6183.15	-6182.45	-6065.08

Table 5. Final log-likelihood of the models estimated for estimation subsets of renters data

Subset	Number of Observations	Initial LL	Final LL		
			CMNL	rCMNL	ICMNL
Subset1	305	-1894.23	-1344.86	-1330.40	-1330.09
Subset2	305	-1894.23	-1343.48	-1343.92	-1330.08
Subset3	306	-1900.44	-1338.28	-1337.87	-1325.61
Subset4	306	-1900.44	-1334.37	-1334.02	-1321.36
Subset5	306	-1900.44	-1342.02	-1341.71	-1331.67

The five validation subsets (20% of the sample) are then used to validate the estimated model outcomes. The predictive power of each of the model is evaluated using both disaggregate level measures of fit (predictive rho-square and average probability of correct prediction) and aggregate level measures of fit (root mean square error and mean absolute deviation between predicted and actual share). Predictive measures of fit for all the models in different subsets are computed and summarized in Table 6 (owners subset) and Table 7 (renters subset) where the improvements in percentage over the CMNL model are presented in the parenthesis.

For owners, the ICMNL model shows improved performance over the CMNL and rCMNL models in all subsets in terms of all measures of fit. However, the performance of the rCMNL model is same as the CMNL model performance in most of the subsets and marginally better in some cases in term of all measures of fit.

For renters, the ICMNL model performs better than the CMNL and the rCMNL models in all validation subsets in terms of the average probability of correct prediction and four out of five subsets (except subset 1) in terms of predicted rho-square. The ICMNL model performs worse than the CMNL and the rCMNL models in terms of root means square error and mean absolute deviation between actual and predicted share in one out of five subsets (subset S4). This is also likely due to the fact that this specific subset may contain a high concentration of observations where households have a lower preference for the alternatives close to their current homes.

Table 6. Ownership model measures of fit in validation subsets

Validation Tools	Subsets	CMNL	rCMNL	ICMNL
Average probability of correct prediction	S1	0.026	0.027	0.030
		(0)	(3.2)	(16.4)
	S2	0.027	0.027	0.032
		(0)	(0)	(19)
	S3	0.027	0.028	0.032
		(0)	(3.7)	(17.6)
	S4	0.028	0.028	0.033
		(0)	(0)	(17.7)
	S5	0.029	0.029	0.034
		(0)	(0)	(18.7)
Root mean square error between predicted and actual share(RMSE)	S1	1.081	1.08	1.017
		(0)	(-0.1)	(-5.9)
	S2	0.899	0.899	0.857
		(0)	(0)	(-4.6)
	S3	1.047	1.046	1
		(0)	(-0.1)	(-4.5)
	S4	0.725	0.725	0.687
		(0)	(0)	(-5.2)
	S5	1.116	1.116	1.089
		(0)	(0)	(-2.4)
Mean absolute deviation between predicted and actual share (MAD)	S1	0.811	0.81	0.753
		(0)	(-0.1)	(-7.2)
	S2	0.723	0.723	0.695
		(0)	(0)	(-3.9)
	S3	0.826	0.825	0.776
		(0)	(-0.1)	(-6.1)
	S4	0.541	0.541	0.516
		(0)	(0)	(-4.6)
	S5	0.893	0.893	0.875
		(0)	(0)	(-2)
Predicted rho-sq	S1	0.325	0.326	0.335
		(0)	(0.3)	(3.1)
	S2	0.335	0.335	0.35
		(0)	(0)	(4.3)
	S3	0.329	0.329	0.344
		(0)	(0)	(4.4)
	S4	0.34	0.341	0.355
		(0)	(0.3)	(4.5)
	S5	0.333	0.333	0.343
		(0)	(0)	(2.8)

Table 7. Renting model measures of fit in validation subsets

Validation Tools	Subsets	CMNL	rCMNL	ICMNL
Average probability of correct prediction	S1	0.026	0.027	0.030
		(0)	(3.2)	(16.4)
	S2	0.027	0.027	0.032
		(0)	(0)	(19)
	S3	0.027	0.028	0.032
		(0)	(3.7)	(17.6)
	S4	0.028	0.028	0.033
		(0)	(0)	(17.7)
	S5	0.029	0.029	0.034
		(0)	(0)	(18.7)
Root mean square error between predicted and actual share(RMSE)	S1	1.081	1.08	1.017
		(0)	(-0.1)	(-5.9)
	S2	0.899	0.899	0.857
		(0)	(0)	(-4.6)
	S3	1.047	1.046	1
		(0)	(-0.1)	(-4.5)
	S4	0.725	0.725	0.687
		(0)	(0)	(-5.2)
	S5	1.116	1.116	1.089
		(0)	(0)	(-2.4)
Mean absolute deviation between predicted and actual share (MAD)	S1	0.811	0.81	0.753
		(0)	(-0.1)	(-7.2)
	S2	0.723	0.723	0.695
		(0)	(0)	(-3.9)
	S3	0.826	0.825	0.776
		(0)	(-0.1)	(-6.1)
	S4	0.541	0.541	0.516
		(0)	(0)	(-4.6)
	S5	0.893	0.893	0.875
		(0)	(0)	(-2)
Predicted rho-sq	S1	0.325	0.326	0.335
		(0)	(0.3)	(3.1)
	S2	0.335	0.335	0.35
		(0)	(0)	(4.3)
	S3	0.329	0.329	0.344
		(0)	(0)	(4.4)
	S4	0.34	0.341	0.355
		(0)	(0.3)	(4.5)
	S5	0.333	0.333	0.343
		(0)	(0)	(2.8)

7 Conclusions

This study proposes an improvement of the existing CMNL model (called ICMNL model) for behavioural choice set consideration with a better approximation to the classical Manski method. The proposed ICMNL model is evaluated in this study using simulated data and then applied to real-world residential location choice data. In both cases, better performance of the ICMNL model is observed compared to the existing CMNL and rCMNL model. Modelling of residential location choice with implicit choice set consideration also produces a behavioural difference in the choice set consideration of owners and renters.

Although the ICMNL model is found to outperform the CMNL and rCMNL models, it still has avenues for further improvements. The threshold effect considered in the model for utility penalization is exogenous and homogeneous across all respondents. The method can be improved by allowing individual specific threshold or threshold specific to the group of respondents belonging the same characteristics.

Further, the choice sets of all individuals in the proposed ICMNL model are constrained because utilities are penalized if alternatives do not meet the criteria of exogenous constraint. However, some individuals can have unconstrained choice sets in reality. Adopting latent classes in the ICMNL model could be a potential direction for further improvement of the proposed ICMNL model where the choice set of one class can be constrained and another class can be unconstrained. Considering the same attributes in the choice set part (in the penalty term) and in the systematic utility may result in an identification issue. The ICMNL model with latent classes for constrained and unconstrained choices set can avoid this issue. However, this technique cannot be applied in this study due to data limitation and could be a future direction of research.

Although the potential of the proposed method observed in this study to capture individual choice set is promising, more testing is recommended with other data sets as a topic of future research. Testing the validity of the findings in other contexts (e.g., route choice, destination choice, activity choice etc.) can also be an interesting direction for future research.

However, with better behavioural grounding (supported by the better model fit) as well as computational tractability, the proposed ICMNL model can be an attractive option for modelling with a large universal choice set where the classical probabilistic approach is infeasible.

Acknowledgements

The authors thank the Economic and Social Research Council (ESRC) and the Institute for Transport Studies, University of Leeds for funding this study. Professor Stephane Hess acknowledges the support of the European Research Council through the consolidator grant 615596-DECISIONS.

References

Arentze, T., & Timmermans, H. (2005). An analysis of context and constraints-dependent shopping behavior using qualitative decision principles. *Urban Studies, 42*(3), 435–448.

Bell, M. G. (2007). Mixed routing strategies for hazardous materials: Decision-making under complete uncertainty. *International Journal of Sustainable Transportation, 1*(2), 133–142.

Ben-Akiva, M., & Lerman, S. (1974). Some estimation results of a simultaneous model of auto ownership and mode choice to work. *Transportation, 3*, 357–376.

Bhat, C. R. (2015). A comprehensive dwelling unit choice model accommodating psychological constructs within a search strategy for consideration set formation. *Transportation Research Part B: Methodological, 79*, 161–188.

Bhat, C. R., & Guo, J. (2004). A mixed spatially correlated logit model: Formulation and application to residential choice modelling. *Transportation Research Part B: Methodological, 38*(2), 147–168.

Bhat, C. R., & Guo, J. Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological, 41*(5), 506–526.

Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2010). Analysis of implicit choice set generation using a constrained multinomial logit model. *Transportation Research Record: Journal of the Transportation Research Board, 2175*, 92–97.

Caicedo, F., Lopez-Ospina, H., & Pablo-Malagrida, R. (2016). Environmental repercussions of parking demand management strategies using a constrained logit model. *Transportation Research Part D: Transport and Environment, 48*, 125–140.

Cascetta, E., & Papola, A. (2001). Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand. *Transportation Research Part C: Emerging Technologies, 9*(4), 249–263.

Cascetta, E., & Papola, A. (2009). Dominance among alternatives in random utility models. *Transportation Research Part A: Policy and Practice, 43*(2), 170–179.

Castro, M., Martínez, F., & Munizaga, M. A. (2013). Estimation of a constrained multinomial logit model. *Transportation, 40*(3), 563–581.

Farooq, B., & Miller, E. J. (2012). Towards integrated land use and transportation: A dynamic disequilibrium-based microsimulation framework for built space markets. *Transportation Research Part A: Policy and Practice, 46*(7), 1030–1053.

Guevara, C. A. (2010). *Endogeneity and sampling of alternatives in spatial choice models.* (Unpublished doctoral dissertation) Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Habib, M., & Miller, E. (2009). Reference-dependent residential location choice model within a relocation context. *Transportation Research Record: Journal of the Transportation Research Board, 2133*, 92–99.

Haque, M. B., Choudhury, C., & Hess, S. (2018). *Investigating the temporal dynamics of long-term and medium-term residential location choices: A case-study of London*. Presented at the Transportation Research Board Annual Meeting, Washington, DC.

Kaplan, S., Bekhor, S., & Shiftan, Y. (2011). Development and estimation of a semi-compensatory residential choice model based on explicit choice protocols. *The Annals of Regional Science, 47*(1), 51–80.

Kwan, M. P., & Hong, X. D. (1998). Network-based constraints-oriented choice set formation using GIS. *Geographical Systems, 5*, 139–162.

Lee, B. H., & Waddell, P. (2010). Residential mobility and location choice: A nested logit model with sampling of alternatives. *Transportation, 37*(4), 587–601.

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision, 8*(3), 22–254.

Martínez, F., Aguila, F., & Hurtubia, R. (2009). The constrained multinomial logit: A semi-compensatory choice model. *Transportation Research Part B: Methodological, 43*(3), 365–377.

Martínez, F., & Hurtubia, R. (2006). Dynamic model for the simulation of equilibrium status in the land use market. *Networks and Spatial Economics, 6*(1), 55–73.

Mcfadden, D. (1978). Modelling the choice of residential location. *Transportation Research Record, 673*, 72–77.

Næss, P. (2009). Residential self-selection and appropriate control variables in land use: Travel studies. *Transport Reviews, 29(*3), 293–324.

Paleti, R. (2015). Implicit choice set generation in discrete choice models: Application to household auto ownership decisions. *Transportation Research Part B: Methodological, 80*, 132–149.

Rashidi, T. H., Auld, J., & Mohammadian, A. (2012). A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection. *Transportation Research Part A: Policy and Practice, 46*(7), 1097–1107.

Scott, D. M. (2006). *Constrained destination choice set generation: Comparison of GIS-based approaches.* Presented at the Transportation Research Board 85th Annual Meeting, Washington, DC.

Swait, J. (2001). A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B: Methodological, 35*(10), 903–928.

Swait, J., & Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological, 21*(2), 91–102.

Termansen, M., McClean, C., & Skov-Petersen, H. (2004). Recreational site choice modelling using high-resolution spatial data. *Environment and Planning B: Planning and Design, 36*, 1085–1099.

Zolfaghari, A. (2013). *Methodological and empirical challenges in modelling residential location choices* (Doctoral thesis). Center for Transport Studies, Imperial College, London.