RESIDENTIAL LOCATION CHOICE BEHAVIOR FOR DIFFERENT HOUSEHOLDS: METHODOLOGY AND CASE STUDY

Pengpeng JIAO
Ph.D.
Department of Urban Engineering
University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo
113-8656 Japan
Fax: +81-3-5841-8527
E-mail: jiao_pengpeng@ut.t.u-tokyo.ac.jp

Noboru HARATA
Professor
Department of Urban Engineering
University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo
113-8656 Japan
Fax: +81-3-5841-8527
E-mail: nhara@ut.t.u-tokyo.ac.jp

Abstract: This paper presents a mixed logit framework to identify residential location choice behavior in households. The model integrates a “direct parametric representation” approach to capture the correlation between spatial units, as well as a comprehensive structure of zonal accessibility to reflect the effects of employment, school, shopping and recreational opportunities. Households are clustered based on demographic and daily trip data to extract their different residential choice characteristics. The model is applied to the central city of Dalian, China. The empirical results reveal that 11675 households are clustered into 5 groups, with distinct characteristics in each group. Results also show the significant differences in sensitivity to female, male and children commuting behavior while households make residential choices, as well as their preferences to zonal accessibility to different activity opportunities.

Key Words: Residential location choice, Mixed logit model, Spatial correlation

1. INTRODUCTION

The integrated land-use and transportation models have been intensively studied in recent years. In this context, one of the most important aspects is residential location choice. It not only determines the household’s activity travel participation, but also influences the association between residential land use and the rest of the urban elements. According to the personal trip survey data of 8 Chinese cities, averagely more than 80 percent of personal daily trips take residential location as origin or destination. As a fast developing country, residential location choice behavior is very important for Chinese urban and transportation planning.

As an appealing tool, discrete choice approach has been broadly applied to residential location choice analysis. Sermons and Koppelman (2001) showed that there are at least two advantages for discrete choice approach to be used in residential location choice problem. First, the discrete residential location choice approach is based on microeconomic random utility theory, and households make the residential location choice as a tradeoff among various locational attributes, such as commute time, house price and access to population and employment. Second, it can include in the utility function other many variables that influence residential location, e.g., socio-demographic characteristics.

Waddell (1996), Tu and Goldfinch (1996), Freeman and Kern (1997), Abraham and Hunt (1997), Ben-Akiva and Bowman (1998), Sermons and Koppelman (1998), Sermons (2000), and Sermons and Koppelman (2001) have further developed this work, and studied households’ residential location choice behavior, as well as the large mobility-travel decision-making framework, with residential choice as one element. More recently, Bhat and Guo (2004), Guo and Bhat (2004) and Miyamoto et al. (2004) have also formulated different residential choice models respectively. Bhat and Guo (2004) proposed a mixed spatially correlated logit model, and represented a powerful approach to capture both random taste variations and spatial correlation. In their work the spatial correlation was modeled with an arbitrary spatial allocation rule. Guo and Bhat (2004) put forward a multi-scale modeling structure, and represented the notion of a neighborhood being a hierarchy of residential groupings. Miyamoto et al. (2004) formulated a mixed logit model with the autoregressive deterministic and error terms, which represented the spatial correlation with a spatial stochastic process.

With some studies focusing on location choice for specific demographic groups (such as single worker or two-worker households), most of above works have analyzed the residential choice behavior for all households, and few of them have distinguished the different characteristics of different household clusters. In fact, different households should have specific behavior characteristics according to the family structure, e.g., households with only workers maybe concentrate most on commute time from home to workplace, while households with school children will pay more attention to the level of schools near to their home, as well as the commute time from home to school. Furthermore, even in the same kind of household, the characteristics of male, female and children should be different. The analysis of such characteristics should be an interesting topic. Sermons and Koppelman (2001) represented the differences between different households, and testified the household responsibility hypothesis, i.e., women commute less than men because women tend to perform more of the household maintenance and child-rearing responsibilities. However, they only focused on commute behavior, the analysis was based on only those households with one female and one male worker, and also they did not investigate the difference between children and adults.

Many researchers have proposed the household classification criteria, e.g., Jones et al. (1983) classified households into eight groups: a) Younger (married) adults without children; b) Families with pre-school children; c) Families with pre-school children and young school children; d) Families with young school children; e) Families with older school children; f) Families of adults, all of working age; g) Older adults, no children in household; h) Retired persons. This clustering method can be applied in developed countries. However, it is not appropriate to Chinese situation. Since most Chinese families have only one child, group c in above criteria will be no sense, and also there will be much less households in group b, d and e than other groups. The cluster is not evenly distributed. Therefore, households must be clustered according to Chinese specific situation.

For the explanatory variables, previous researches have made use of rather broad information, including socio-demographic data, economic data, environmental data, accessibility data and housing specific data (Guo and Bhat, 2004). All of them provided a very rich set of variables for consideration in model specification. As a bridge between transportation system and urban land use structure, accessibility of spatial unit is a rather comprehensive and macroscopic index, therefore it is powerful for analysis based on different household clusters, especially for developing countries. Bhat and Guo (2004) used the Hansen-type accessibility proposed
by Fotheringham (1986), however, it is rather straightforward, and not very appropriate for behavioral analysis. Moreover, the accessibility to school opportunity should also be considered in the analysis for different household clusters.

One key feature of this paper is to cluster the households based on Chinese situation, and to analyze the residential location choice behavior according to different household clusters. Another key feature is to further revise the spatial correlation in the discrete choice model structure, as well as to reflect the zonal accessibility for employment, school, shopping and recreational opportunities. And the third key feature, to present the differences among male, female and children in residential choice behavior.

The rest of this paper is organized as follows. Methodology is proposed in Section 2, including a mixed logit model specification of the residential choice, the spatial correlation formulation, the revised accessibility structure, the household cluster approach, and some issues in estimation. The case study area of Dalian city, China is described in Section 3, as well as data sources. Estimation results of the mixed residential location choice model are reported in Section 4, along with the household clustering results. Conclusions and further researches are summarized in the last section.

2. METHODOLOGY

As discussed above, residential location is discrete in nature. Meanwhile, to incorporate the spatial correlation between units and random taste variations of households, a specification more flexible than the MNL model is required. This leads to the choice of mixed logit model (Train, 2003), in which the parameters are assumed to be random across households to reflect the taste variations, and spatial correlation is incorporated in the random term. The following sections describe the model specification and some related considerations, including spatial correlation formulation, accessibility representation, household cluster approach and some issues in estimation.

2.1 Mixed Logit Model Specification

The residential location choice is decided by households in nature. Therefore, random utility theory can be applied here, in which the residential location is determined by the following equation:

\[ y_{qi} = \begin{cases} 1 & \text{if } U_{qi} \geq U_{qj} \text{ for } j = 1, \ldots, J \\ 0 & \text{Otherwise} \end{cases} \]  

where \( y_{qi} \) is an indicator for household \( q \) to select spatial unit \( i \); \( U_{qi} \) is the utility function; \( J \) is the total number of alternatives, i.e., the number of spatial units.

As usual, the utility function \( U_{qi} \) can be written in two terms: a deterministic term \( V_{qi} \) and a random term \( \varepsilon_{qi} \):

\[ U_{qi} = V_{qi} + \varepsilon_{qi} \]  

In order to allow for spatial correlations between residential units, the random term \( \varepsilon_{qi} \) is written in the following formulation:

\[ \varepsilon_{qi} = \xi_{qi} + \zeta_{qi} \]
where $\xi_{qi}$ is a random factor with zero mean, $\xi_{qi} = \sigma_i \pi_{qi}$; $\sigma_i$ is parameter to be estimated; $\pi_{qi} \sim N(0, \Sigma)$ is a standard multivariate normal distribution; $\Sigma$ is the covariance matrix of $\pi_{qi}$ reflecting the spatial correlation, and it can be obtained using approach represented in the following section; $\xi_{qi}$ is assumed to be iid Gumbel distribution.

Based on above formulations, the choice probability for household $q$ to select spatial unit $i$ is:

$$L_{qi} = \frac{\exp(V_{qi} + \sigma_i \pi_{qi})}{\sum_{j=1}^{J} \exp(V_{qj} + \sigma_j \pi_{qj})}$$

(4)

The deterministic term of the utility function, $V_{qi}$, can be written as following:

$$V_{qi} = \sum_{k=1}^{K} \beta_k x_{qik}$$

(5)

where $\beta_k$ is parameter to be estimated; $x_{qik}$ is explanatory variables, e.g., spatial unit information, household information, etc.; $K$ is the number of explanatory variables.

To incorporate random taste variations, parameters $\beta_k$ are assumed to be multivariate normal over households. Let $f$ represent the density function of the multivariate normal distribution, the unconditional choice probability is therefore the integral of $L_{qi}$ over all possible variables of $\beta_k$:

$$P_{qi} = \int_{-\infty}^{\infty} \frac{\exp(\sum \beta_k x_{qik} + \xi_{qi})}{\sum_{j=1}^{J} \exp(\sum \beta_k x_{qj} + \xi_{qj})} \cdot f(\beta) d\beta$$

(6)

In this research, based on data availability and residential choice mechanism, the explanatory variables include the absolute difference between household car ownership and zonal average car ownership, male commute time, female commute time, children school commute time, zonal area, population density, and accessibility to employment, school, shopping and recreational opportunities.

2.2 Spatial Correlation

Anselin (1988, 1999) and Wang and Kockelman (2005) suggested that there are three principal methods for addressing the spatial effects that exist in land use: a specification of a spatial stochastic process, direct parametric representation of the covariance structure, and a non-parametric framework. Spatial autoregressive (SAR) in Miyamoto et al (2004) belongs to the spatial stochastic process, which is most often used. It formulates a functional relationship between a random variable at a given location and this same random variable at other locations. The variance structure then follows from the nature of the stochastic process. One frequently encountered problem is the lack of stationarity in this process, which has important implications for the types of central limit theorems and laws of large numbers that need to be invoked to obtain asymptotic properties for estimators and specification test (Anselin, 2006). Spatial correlation in Wang and Kockelman (2005) is a kind of “direct parametric
representation”. It expresses the elements of the covariance matrix in a parsimonious fashion as a “direct” function of distances. In this way, it will not induce heteroskedasticity, and always meet stationary requirements. A key issue in this method is that it is fraught with a number of estimation problems because of flexible expressions (Anselin, 2006). In the third approach, nonparametric framework, the spatial covariance is estimated from the sample covariance for the residuals of each set of location pairs, and it requires panel data, with the time dimension much greater than the cross-sectional dimension (Anselin, 2006). This approach is seldom used because of lack of data.

In this research, the “direct parametric representation” is used to avoid the covariance-stationary problem. Moreover, it can reflect the relation between spatial correlation and distance more effectively. Since the spatial correlation should decrease with the increase of distance, two candidates, negative exponential and inverse square distance, are adopted here, as follows.

\[
\Sigma_{ij} = 1/(d_{ij})^\alpha \\
\Sigma_{ij} = \exp(-d_{ij}^\gamma)
\]

where \(d_{ij}\) is the distance between spatial unit \(i\) and \(j\); \(\alpha\) and \(\gamma\) are parameters to be estimated, which can be determined according to the likelihood ratio in the estimation results.

### 2.3 Accessibility

As a bridge between transportation system and land use pattern, accessibility is a rather powerful explanatory variable for residential location choice analysis. Geertman and Ritsema (1995), Vickerman (1974) and Ghtierrez (2001) studied on accessibility respectively. According to their works, accessibility can be classified into three types: average weighted travel time accessibility, economic potential accessibility and daily accessibility. The first type is the measurement of travel time between one spatial unit to the others, which is closely related to the location, economic status and transportation infrastructure level of the unit. The second type of accessibility is determined by the economic potential of the unit, which can be the GDP, population or gross sales of consumer goods. The daily accessibility reflects the number of activities from one spatial unit to others in one day, which can be obtained by survey. Since this research mainly focuses on planning purpose, the economic potential accessibility is adopted here.

The original formulation of economic potential accessibility is in the following formulation:

\[
A_i = \sum_{j=1}^{n} \frac{M_j}{R_{ij}^\theta} 
\]

where \(A_i\) is the accessibility index of spatial unit \(i\); \(M_j\) is the economic potential of spatial unit \(j\), which can be GDP, population or gross sales of consumer goods; \(R_{ij}\) is the generalized impedance between \(i\) and \(j\), which can be distance, travel time or travel cost; \(\theta\) is the impedance coefficient.

To make it more understandable and more appropriate for residential choice analysis, the above formulation is revised as follows.

The accessibility of spatial unit \(i\) between \(i\) and \(j\) is:

\[
A_{ij} = \Gamma \cdot M_j \cdot f(R_{ij})
\]
where $A_{ij}^l$ is the accessibility of spatial unit $i$ between $i$ and $j$; $\Gamma$ is a coefficient; $f(R_{ij})$ is the distribution impedance function. Assume the sum of accessibility of all spatial units is one, we get:

$$\sum_i \sum_j \Gamma \cdot M_{ij} \cdot f(R_{ij}) = 1 \quad (11)$$

Therefore,

$$\Gamma = \frac{1}{\sum_i \sum_j M_{ij} \cdot f(R_{ij})} \quad (12)$$

Replace equation (12) in (10), one can get:

$$A_{ij}^l = \frac{M_{ij} \cdot f(R_{ij})}{\sum_i \sum_j M_{ij} \cdot f(R_{ij})} \quad (13)$$

Furthermore, the accessibility of spatial unit $i$ is:

$$A_i = \frac{\sum_j M_{ij} \cdot f(R_{ij})}{\sum_i \sum_j M_{ij} \cdot f(R_{ij})} \quad (14)$$

From this formulation, we can obtain the accessibility of any spatial unit. Here $f(R_{ij})$ is in the form of equation (7) and (8), which can also be determined according to the likelihood ratio in the estimation results.

In this research, we utilized distance, i.e., $d_{ij}$, as the impedance variable between two spatial units. Therefore, one can get the formulations of accessibility to different opportunities as follows:

$$A_i^{Sch} = \frac{\sum_j S_j^{Edu} \cdot f(d_{ij})}{\sum_i \sum_j S_j^{Edu} \cdot f(d_{ij})} \quad (15)$$

$$A_i^{Emp} = \frac{\sum_j N_j^{Emp} \cdot f(d_{ij})}{\sum_i \sum_j N_j^{Emp} \cdot f(d_{ij})} \quad (16)$$

$$A_i^{Shop} = \frac{\sum_j G_j^{Shop} \cdot f(d_{ij})}{\sum_i \sum_j G_j^{Shop} \cdot f(d_{ij})} \quad (17)$$

$$A_i^{Rec} = \frac{\sum_j S_j^{Rec} \cdot f(d_{ij})}{\sum_i \sum_j S_j^{Rec} \cdot f(d_{ij})} \quad (18)$$

where $S_j^{Edu}$ is the area of education land use in spatial unit $j$; $N_j^{Emp}$ is the number of employment opportunities in spatial unit $j$; $G_j^{Shop}$ is the gross sales of consumer goods in spatial unit $j$; $S_j^{Rec}$ is the area of recreational land use in spatial unit $j$.

### 2.4 Household Cluster

As stated above, different households have different residential choice characteristics.
Furthermore, Chinese household classification should be greatly different from western developed countries. Therefore, new approach different from those in the literature must be developed according to Chinese situation.

In this research, a large amount of information about the households is collected through survey, and eleven groups of data are extracted to represent the household characteristics, including demographic data, household trip data, etc. Since it is difficult for cluster analysis with too much kind of data, factor analysis is first carried out to simplify the data. Here the principal component analysis (PCA) is used for data reduction. Based on the values of principal components obtained, cluster analysis is then implemented to cluster the households. The detailed data and clustering method are described in the results.

2.5 Some Issues in Estimation
The parameters to be estimated in the mixed logit model include $\sigma$ representing spatial correlation and multivariate normally distributed $\beta$ characterizing random taste variations. The log-likelihood function is:

$$L(\beta, \sigma) = \sum_{q} \sum_{y_{q}} y_{q} \ln P_{q}(\beta, \sigma)$$

(19)

where $\beta$ and $\sigma$ are independent with each other.

Estimation can be implemented with maximum simulated likelihood (MSL) method (Bhat and Guo, 2004; Train, 2003), which has been frequently used.

To reflect the correlation between spatial units, one must generate multivariate normally distributed random draws with the correlation matrix in equation (7) and (8). Usually the multivariate normal distribution is generated using a Cholesky decomposition, which requires the covariance matrix to be symmetric and positive definite (Train, 2003). However, the spatial correlation matrix can only be guaranteed to be symmetric, not positive definite, and the Cholesky decomposition can not be implemented here. In this research, generalized Cholesky decomposition in Gill and King (2004) was adopted, which can also be found in Wang and Kockelman (2005). Using the pseudo triangular matrices generated by generalized Cholesky decomposition, one just needs to generate iid normally distributed random draws.

Random draws are generated using randomly scrambled Halton method in Bhat (2003), which has been demonstrated to be more uniformly distributed and efficient than pseudo Monte Carlo (PMC) method.

All estimations in this research were implemented using the GAUSS programming language. Bhat’s (2003) GAUSS code for scrambled Halton sequence was modified and integrated into our MSL estimation code.

3. CASE STUDY AREA AND DATA
The case study area selected for this study is central city of Dalian, China, which is located in the center of Northeastern Asian economic region. Among the three major economic zones in China, Dalian is one of the central city in the Ring-Bohai economic zone. And in Liaoning Province, Dalian is the leading city in stimulating the economic development of the whole province and also the largest port in Northeast and Inner Mongolia region. As shown in Figure 1, Dalian is serving as a window from China to the world. The problem is residential location
choice in the 35 spatial units in Dalian. Here the traffic analysis zones (TAZ) are taken as the alternatives for residential location choice.

The primary source of data is the 2004 Dalian metropolitan area personal trip survey data. This survey collected information about all travel activities undertaken by members of 15544 households on December 13, 2004, as well as the residential locations of households. The survey also obtained individual and household sociodemographic information. Information of 11675 households in the central city was selected from the database for the case study.

In addition to the personal trip survey data, four other data sets associated with Dalian city were used: land use data, demographic data, census data and the zone-to-zone travel LOS data. The land use data was obtained from Dalian Bureau of Urban Planning, and was used to get the total acreage and acreage in specific land use purposes (including residential land, industrial land, shopping land, education land, office land, transportation land, and recreation land, etc.) for each TAZ. The demographic data and census data came from Dalian Municipal Bureau of Statistics, and were used to compute the number of households, the total population, the population density, and gross sales of consumer goods for each TAZ. The zone-to-zone travel LOS data was collected from the 2004 Dalian transportation survey. It provided information on travel between each pair of the 35 zones, and also contained the inter-zonal distances as well as peak and off-peak travel times and costs. Moreover, the commute time for male, female and children are extracted from the personal trip survey data to investigate the differences among them. All above data was used to develop measures of accessibility to school, employment, shopping and recreational opportunities according to equations (15) to (18).

Finally, the following three groups of data were obtained:
1) TAZ related data, including total acreage, acreage in specific land use purposes, total population, population density, number of households, average household size, number of employment opportunities, and accessibility to different activity opportunities for each zone.

2) Household related data, including number of members, number of adults, number of students, age of household head, career of household head, car ownership, average commute trip time, average non-commute trip time, commute trip mode, non-commute trip mode, commute time to employment of male and female workers, and commute time to school of children for each household.

3) Zonal heterogeneity related data, including the absolute difference between zone-average household size and the number of family members in each household, and the absolute difference between zone-average car ownership and household car ownership.

Above data provided a rich set of variables for consideration in model specification.

4. RESULTS

4.1 Household Clustering Results

4.1.1 Factor Analysis

As discussed above, PCA and cluster analysis are first implemented for household clustering. Eleven groups of data of 11675 households are collected to represent the household characteristics, including number of members, number of adults, number of students, age of household head, career of household head, car ownership, number of non-commute trips within one day, average commute time, average non-commute time, commute mode and non-commute mode. One can find out that there are both demographic data and trip survey data, therefore they can reflect the comprehensive characteristics of different household clusters. These data were input into SPSS software for PCA analysis, and varimax orthogonal rotation was used. The KMO measure of sampling adequacy was 0.863, which shows that the data meet the need of factor analysis. The component matrix and rotated component matrix are shown in Table 1 and Table 2 respectively.

From Table 1 and 2 one can find out that the sum of squared loadings is 75.123%, i.e., 75.123 percent of the original information are preserved in the five principal components after the PCA analysis. According to the rotated component matrix, five principal components are extracted from original eleven variables, as follows:

- PCA 1: including number of members, number of adults and number of students, we name it “Household Scale”;
- PCA 2: including car ownership, commute mode and non-commute mode, we name it “Traffic Mode”;
- PCA 3: including career of household head and number of non-commute trips, we name it “Social Communication”;
- PCA 4: including commute time and non-commute time, we name it “Traffic Accessibility”;
- PCA 5: including age of household head, we name it “Life Cycle”.


Table 1 Component matrix of PCA

<table>
<thead>
<tr>
<th>Component</th>
<th>Household Scale</th>
<th>Traffic Mode</th>
<th>Social Communication</th>
<th>Traffic Accessibility</th>
<th>Lifecycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Members</td>
<td>0.799</td>
<td>-0.379</td>
<td>0.326</td>
<td>-0.016</td>
<td>-0.078</td>
</tr>
<tr>
<td>Number of Adults</td>
<td>0.801</td>
<td>-0.406</td>
<td>0.347</td>
<td>-0.024</td>
<td>-0.080</td>
</tr>
<tr>
<td>Number of Students</td>
<td>0.597</td>
<td>-0.344</td>
<td>0.007</td>
<td>0.025</td>
<td>0.097</td>
</tr>
<tr>
<td>Age of Household Head</td>
<td>-0.409</td>
<td>-0.152</td>
<td>-0.095</td>
<td>-0.077</td>
<td>0.620</td>
</tr>
<tr>
<td>Career of Household Head</td>
<td>-0.276</td>
<td>-0.139</td>
<td>0.659</td>
<td>0.075</td>
<td>0.141</td>
</tr>
<tr>
<td>Car Ownership</td>
<td>0.159</td>
<td>0.699</td>
<td>0.089</td>
<td>0.628</td>
<td>0.203</td>
</tr>
<tr>
<td>Number of Non-Commute Trips</td>
<td>-0.497</td>
<td>0.087</td>
<td>0.556</td>
<td>0.018</td>
<td>-0.036</td>
</tr>
<tr>
<td>Commute Time</td>
<td>0.296</td>
<td>0.564</td>
<td>0.081</td>
<td>-0.589</td>
<td>0.205</td>
</tr>
<tr>
<td>Non-Commute Time</td>
<td>0.242</td>
<td>0.606</td>
<td>0.213</td>
<td>-0.512</td>
<td>0.249</td>
</tr>
<tr>
<td>Commute Mode</td>
<td>0.308</td>
<td>0.732</td>
<td>0.087</td>
<td>0.356</td>
<td>-0.267</td>
</tr>
<tr>
<td>Non-Commute Mode</td>
<td>0.296</td>
<td>0.694</td>
<td>0.205</td>
<td>0.352</td>
<td>-0.336</td>
</tr>
<tr>
<td>Sum of % of Variance</td>
<td>22.549</td>
<td>20.193</td>
<td>13.319</td>
<td>11.523</td>
<td>7.539</td>
</tr>
<tr>
<td>Squared Loadings</td>
<td>22.549</td>
<td>42.742</td>
<td>56.061</td>
<td>67.584</td>
<td>75.123</td>
</tr>
</tbody>
</table>

Table 2 Rotated component matrix of PCA

<table>
<thead>
<tr>
<th>Component</th>
<th>Household Scale</th>
<th>Traffic Mode</th>
<th>Social Communication</th>
<th>Traffic Accessibility</th>
<th>Lifecycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Members</td>
<td>0.944</td>
<td>0.050</td>
<td>-0.015</td>
<td>0.024</td>
<td>-0.024</td>
</tr>
<tr>
<td>Number of Adults</td>
<td>0.965</td>
<td>0.034</td>
<td>0.006</td>
<td>0.018</td>
<td>-0.032</td>
</tr>
<tr>
<td>Number of Students</td>
<td>0.638</td>
<td>-0.108</td>
<td>-0.231</td>
<td>-0.032</td>
<td>0.106</td>
</tr>
<tr>
<td>Age of Household Head</td>
<td>-0.047</td>
<td>-0.087</td>
<td>-0.143</td>
<td>-0.053</td>
<td>0.746</td>
</tr>
<tr>
<td>Career of Household Head</td>
<td>0.057</td>
<td>-0.076</td>
<td>0.723</td>
<td>-0.031</td>
<td>0.151</td>
</tr>
<tr>
<td>Car Ownership</td>
<td>0.032</td>
<td>0.965</td>
<td>-0.006</td>
<td>-0.025</td>
<td>0.153</td>
</tr>
<tr>
<td>Number of Non-Commute Trips</td>
<td>-0.253</td>
<td>0.056</td>
<td>0.705</td>
<td>0.002</td>
<td>-0.024</td>
</tr>
<tr>
<td>Commute Time</td>
<td>0.014</td>
<td>0.114</td>
<td>-0.108</td>
<td>0.879</td>
<td>-0.056</td>
</tr>
<tr>
<td>Non-Commute Time</td>
<td>-0.006</td>
<td>0.167</td>
<td>0.030</td>
<td>0.875</td>
<td>0.031</td>
</tr>
<tr>
<td>Commute Mode</td>
<td>-0.037</td>
<td>0.885</td>
<td>-0.115</td>
<td>0.160</td>
<td>0.115</td>
</tr>
<tr>
<td>Non-Commute Mode</td>
<td>0.014</td>
<td>0.909</td>
<td>0.000</td>
<td>0.136</td>
<td>0.059</td>
</tr>
<tr>
<td>Sum of % of variance squared Loadings</td>
<td>20.931</td>
<td>15.500</td>
<td>15.064</td>
<td>14.451</td>
<td>9.177</td>
</tr>
<tr>
<td>Cumulative %</td>
<td>20.931</td>
<td>36.431</td>
<td>51.495</td>
<td>65.946</td>
<td>75.123</td>
</tr>
</tbody>
</table>

Above five principal components reflect the household characteristics comprehensively, and can be used as the basis for cluster analysis.

4.1.2 Cluster Analysis
Cluster analysis was implemented based on the values of five principal components computed from the component matrix. Two-step cluster method was performed using SPSS software. Original 11675 households were clustered into five groups, and the average values of original explanatory variables are shown in Table 3, as well as the number of households in each group. The ranks from high to low of the explanatory variables within different household clusters are also shown in the parentheses. Since the career of household head, commute mode and non-commute mode are categorical data, their average values are no sense, and are
neglected from the table.

<table>
<thead>
<tr>
<th>Table 3 Household cluster results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Members</td>
</tr>
<tr>
<td>--------------------</td>
</tr>
<tr>
<td>Number of Adults</td>
</tr>
<tr>
<td>Number of Students</td>
</tr>
<tr>
<td>Age of Household Head</td>
</tr>
<tr>
<td>Car Ownership</td>
</tr>
<tr>
<td>Non-Commute Trips</td>
</tr>
<tr>
<td>Commute Time (min)</td>
</tr>
<tr>
<td>Number of Households</td>
</tr>
</tbody>
</table>

From Table 3, one can find out the characteristics of each group as follows:

- Group 1: Just married young couples, almost no student, youngest household head, rather high car ownership, less non-commute trips because of the high burden of life, and very long commute and non-commute time. We name it “Young Couple”.
- Group 2: Married couple living together with old parents, about 3 members, no student, young household head, rather lower car ownership, more non-commute trips because of retired old people, and very long commute and non-commute time. We name it “Young Couple + Old”.
- Group 3: Old couple living alone, no student, oldest household head, lowest car ownership because of consumption concept, most number of non-commute trips because of retired people, very short commute and non-commute time because of their activity space. We name it “Old Couple”.
- Group 4: Parents living together with children, also (less) grandparents, most students, middle aged household head, highest car ownership, least non-commute trips because of the responsibility to rear children, shortest commute and non-commute time because of presence of students. We name it “Student Household”.
- Group 5: Grandparents, parents and children living together, most members, middle aged household head, and all other variables rank the 3rd within five groups. We name it “Big Household”.

From above one can find out that every cluster has distinct characteristics. Furthermore, the households are clustered rather evenly, which is shown in the results of the last row. Therefore, this classification method is appropriate to Chinese situation, and can be used as basis for residential choice behavior analysis.

4.2 Estimation Results

The mean values of estimates for five household clusters are presented in Table 4 respectively, with the \( t \)-statistics shown in parentheses. Standard deviations are neglected here because of too much data. Ten explanatory variables are finally decided after several experiments. Children commute time is omitted in the first three groups because there is no student in these clusters. Accessibility to school opportunity is considered in the first two clusters because they will have children in the future, but omitted in the “Old Couple” household.
$1/(d_y)^2$ was finally chosen to compute the spatial correlation matrix because it yielded the highest likelihood ratio among $1/d_y$, $1/d_y^5$, $1/d_y^2$, $1/d_y^3$, $\exp(-d_y)$ and $\exp(-d_y^2)$. Meanwhile, $1/d_y^3$ was finally determined for calculation of accessibility in equations (15) to (18) based on the same criteria.

Every parameter has the sign as expected. The absolute difference between household car ownership and zonal average car ownership has the expected negative sign, indicating that similar households tend to locate close to each other. This collocation appears because housing developers usually supply similar houses in a certain area, i.e., economical apartments are rarely found in the same neighborhood as luxurious houses. The male, female and children commute time have the expected negative signs, i.e., proximity to the employment location or school is an important factor in residential location choice. Zonal area has the expected positive sign, that is, households are more likely to locate in larger zones than smaller zones. The positive sign of zonal population density shows that households are also more likely to locate in zones with high population density, which reflects the population clustering. All four accessibility indices have expected positive signs, indicating that households tend to live in zones with good accessibility to activity opportunities.

Table 4 Results of parameter estimation

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Young Couple</th>
<th>Young Couple + Old</th>
<th>Old Couple</th>
<th>Student Household</th>
<th>Big Household</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute Difference between Household and Zonal Average Car Ownership</td>
<td>-0.735 (-1.02)</td>
<td>-0.937 (-1.06)</td>
<td>-1.518 (-0.84)</td>
<td>-1.133 (-1.12)</td>
<td>-1.249 (-1.58)</td>
</tr>
<tr>
<td>Male Commute Time (min)</td>
<td>-0.061 (-2.91)</td>
<td>-0.063 (-2.75)</td>
<td>-0.046 (-1.69)</td>
<td>-0.036 (-2.61)</td>
<td>-0.034 (-2.31)</td>
</tr>
<tr>
<td>Female Commute Time (min)</td>
<td>-0.069 (-2.16)</td>
<td>-0.072 (-2.07)</td>
<td>-0.038 (-1.52)</td>
<td>-0.047 (-2.02)</td>
<td>-0.052 (-1.97)</td>
</tr>
<tr>
<td>Children Commute Time (min)</td>
<td>\</td>
<td>\</td>
<td>\</td>
<td>-0.063 (-3.10)</td>
<td>-0.059 (-2.79)</td>
</tr>
<tr>
<td>Area of the Spatial Unit Selected (km²)</td>
<td>0.193 (2.05)</td>
<td>0.216 (2.09)</td>
<td>0.283 (2.47)</td>
<td>0.237 (2.19)</td>
<td>0.261 (2.37)</td>
</tr>
<tr>
<td>Population Density (10 thousand persons per km²)</td>
<td>1.829 (1.72)</td>
<td>2.693 (2.35)</td>
<td>3.385 (2.83)</td>
<td>2.815 (2.22)</td>
<td>3.101 (2.41)</td>
</tr>
<tr>
<td>Accessibility to Employment Opportunity</td>
<td>2.914 (2.13)</td>
<td>3.436 (2.03)</td>
<td>1.018 (1.72)</td>
<td>2.015 (1.83)</td>
<td>2.132 (1.92)</td>
</tr>
<tr>
<td>Accessibility to School Opportunity</td>
<td>1.728 (1.85)</td>
<td>1.907 (1.73)</td>
<td>\</td>
<td>2.532 (2.09)</td>
<td>2.358 (2.04)</td>
</tr>
<tr>
<td>Accessibility to Shopping Opportunity</td>
<td>1.693 (1.47)</td>
<td>1.822 (1.63)</td>
<td>1.592 (2.16)</td>
<td>1.314 (1.21)</td>
<td>1.615 (1.84)</td>
</tr>
<tr>
<td>Accessibility to Recreational Opportunity</td>
<td>0.538 (1.07)</td>
<td>0.819 (1.39)</td>
<td>1.237 (1.98)</td>
<td>0.739 (1.12)</td>
<td>0.713 (1.81)</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>1692</td>
<td>1845</td>
<td>2705</td>
<td>3821</td>
<td>1612</td>
</tr>
<tr>
<td>Log-likelihood at Convergence</td>
<td>-925.2</td>
<td>-933.2</td>
<td>-944.6</td>
<td>-921.5</td>
<td>-947.4</td>
</tr>
<tr>
<td>Adjusted Log-likelihood Ratio</td>
<td>0.181</td>
<td>0.174</td>
<td>0.164</td>
<td>0.184</td>
<td>0.161</td>
</tr>
</tbody>
</table>

Further observe the parameters of four accessibility indices, one can find out that accessibility to employment opportunities is the most important for “Young Couple” and “Young Couple + Old” households, while accessibility to school opportunities is the most important for “Student” and “Big” households, because these two kinds of households concentrate more on children’s school opportunities when considering residential location choice. For the “Old
Couple” household, the accessibility to shopping and recreational opportunities is more important, because most members in this cluster have already retired.

Table 5 is extracted from Table 4 for more detailed analysis. It is obvious that almost all of the female parameters are greater in magnitude than the male commute time parameters, except for “Old Couple” cluster. These differences are most pronounced for households with children and old people. One interpretation of this phenomenon is the household responsibility hypothesis, i.e., women are more sensitive to commute time because they tend to perform more of the household maintenance and child-rearing responsibilities. It also suggests that male’s value to the household is his income-earning ability, while the female’s value to the home includes both her income-earning ability and her household maintenance skill. The difference between parameters of female and male commute time is biggest in “Big Household”, because in this cluster, female need to rear children, as well as to take care of old people.

Meanwhile, for households with students, the student commute time has the biggest weight when considering residential location choice, because households concentrate more on children. It is appropriate to Chinese current situation.

The exception of “Old Couple” is due to Chinese tradition, i.e., in the past years, male is more powerful than female in the household.

### Table 5 Parameter estimates of commute time

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Young Couple</th>
<th>Young Couple + Old</th>
<th>Old Couple</th>
<th>Student Household</th>
<th>Big Household</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Commute Time (min)</td>
<td>-0.061</td>
<td>-0.063</td>
<td>-0.046</td>
<td>-0.036</td>
<td>-0.034</td>
</tr>
<tr>
<td>Female Commute Time (min)</td>
<td>-0.069</td>
<td>-0.072</td>
<td>-0.038</td>
<td>-0.047</td>
<td>-0.052</td>
</tr>
<tr>
<td>Children Commute Time (min)</td>
<td>\</td>
<td>\</td>
<td>\</td>
<td>-0.063</td>
<td>-0.059</td>
</tr>
<tr>
<td>Difference between Male and Female</td>
<td>0.008</td>
<td>0.009</td>
<td>0.008</td>
<td>0.011</td>
<td>0.018</td>
</tr>
<tr>
<td>Commute Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 5. CONCLUSIONS

In this paper, a mixed logit model for the analysis of residential location choice behavior in households has been proposed. The correlation between spatial units is presented in a “direct parametric representation” approach, and zonal accessibility indices to employment, school, shopping and recreational opportunities are formulated in a rather comprehensive structure. Since there should be significant characteristics in residential location choice behavior for different households, a clustering approach is put forward based on Chinese situation. All above issues are integrated into the discrete choice model structure.

The households are classified into several groups based on PCA and cluster analysis. The mixed logit model is estimated through MSL method, with randomly scrambled Halton sequences and generalized Cholesky decomposition technique integrated in the approach. The empirical analysis is implemented to examine the residential choice behavior of households in the central city of Dalian, China, and a number of conclusions can be made from these empirical results.

First, households can be classified into several clusters based on demographic and trip data.
Each cluster has distinct characteristics.

Second, there are significant differences in sensitivity to female, male and children commuting behavior when households make residential location choices. Female workers are more sensitive than male workers, and children are the most sensitive.

Third, difference between female and male commuting behavior is the most pronounced in households with children and old people, which can be explained using household responsibility hypothesis.

Fourth, in “Old Couple” households, male is more sensitive to commute time, which is due to Chinese tradition, i.e., male is more powerful than female in traditional Chinese households.

Fifth, different households pay more attention to different accessibility. “Young Couple” and “Young Couple + Old” households concentrate on accessibility to employment opportunities, while “Student” and “Big” households pay more attention to accessibility to school opportunities. Accessibility to shopping and recreational opportunities is more important for “Old Couple” households.

Except for explanatory variables considered in this paper, household income and land price should also influence residential choice to some extent. These variables are left out of the model because of lack of such data. In addition, the use of panel data to capture the dynamics of household residential location changes looks also interesting. Both are suggested as areas for future researches.

ACKNOWLEDGEMENT

This research is sponsored by the JSPS (Japan Society for the Promotion of Science) fellowship. The authors would like to thank the Institute of Transportation Engineering (ITE) of Tsinghua University for providing the data used in the analysis. The authors are also grateful to the anonymous referees for the careful review of this paper.

REFERENCES


