

**Univariate or Multivariate Analysis for Better Prediction Accuracy? A Case Study of  
Heterogeneity in Vehicle Ownership**

**Sabreena Anowar**

Postdoctoral Associate

Department of Civil, Environmental and Construction Engineering

University of Central Florida

Tel: 1-407-823-4815; Fax: 1-407-823-3315

Email: [sabreena.anowar@ucf.edu](mailto:sabreena.anowar@ucf.edu)

**Naveen Eluru\***

Associate Professor

Department of Civil, Environmental and Construction Engineering

University of Central Florida

Tel: 1-407-823-4815; Fax: 1-407-823-3315

Email: [naveen.eluru@ucf.edu](mailto:naveen.eluru@ucf.edu)

\*Corresponding author

**ABSTRACT**

In this research, we contrast different modeling frameworks that offer alternative ways of capturing observed/unobserved heterogeneity. The model systems compared are: ordered logit, residential location cluster based ordered logit model (exogenous segmentation), mixed ordered logit, latent segmentation based ordered logit model, and a joint copula based self-selection clustering model. While the comparison across single dependent variable models is straight forward, the comparison with the copula based model requires post-processing to generate marginal distribution for the choice of interest. The comparison exercise is conducted in the vehicle ownership context using O-D survey data of Greater Montreal Area (GMA), Canada. The performance of the alternative frameworks is examined in the context of model estimation and validation (at the aggregate and disaggregate level) using a host of comparison metrics. In all cases, the superior performance of the ordered part of the joint copula based model indicates that employing information from an additional dependent variable (such as residential location choice in our case) allows us to better understand and predict the main dimension of interest (vehicle ownership).

**Keywords:** vehicle ownership, copula models, latent class models, residential self-selection, mixed ordered logit, residential clustering, observed heterogeneity, unobserved heterogeneity

## 1. INTRODUCTION

### 1.1 Methodological Context

Discrete choice models form an important analytical tool for analyzing choice behavior in various fields including transportation, bio-statistics, epidemiology, marketing, and health. Specifically, in transportation, several travel behavior related choices (such as travel mode choice, vehicle ownership, travel frequency, residential location, and destination location) are investigated using discrete choice models. Based on the nature of the dependent variable, either an ordered or unordered model structure is employed for model development. Traditional models such as ordered logit (OL)/probit (OP) model (for ordinal variables) and multinomial logit (MNL) model served as the workhorses for model development in the transportation field up to 2000. These models, implicitly restrict the impact of exogenous variables on the choice process to be the same across the entire population – referred to as the *homogeneity* assumption. The assumption, when violated, could potentially result in biased model parameter estimates (see Chamberlain, 1980; Bhat, 2015).

Several approaches have been proposed to relax the homogeneity assumption (Manning et al., 2016). We discuss four approaches relevant to our study context. *First* and probably the most straight forward approach suggested to address the restrictive homogeneity assumption is clustering of the population based on exogenous variables (for instance, transportation infrastructure attributes, built environment measures, and/or travel behavior characteristics) and developing cluster specific models (Salon, 2015; Jacques and El-Geneidy, 2014; Song and Knaap, 2007; Damant-Sirois et al., 2014; Jacques et al., 2013; Bachand-Marleau et al., 2011). However, based on the number of exogenous candidate variables of interest, the number of mutually exclusive sample clusters could increase very rapidly; thus increasing the number of models to be developed (see Eluru et al., 2012a for more discussion). Further, small number of observations in some of the clusters might cause model estimation and interpretational problems.

A *second* approach to address homogeneity is to allow for the impact of exogenous variables to follow a distributional assumption (generally normal, log-normal, triangular, or uniform) as opposed to restricting the impact to a single value across the population. The approach, often referred to as mixed discrete choice models, accommodates for unobserved heterogeneity across the population and improves parameter accuracy. Several research efforts in transportation have employed these models (see Bhat et al., 2008 for a detailed review). In general, these approaches are focused on the unobserved component of the model and usually require extensive simulation for model estimation. In a frequentist approach, Maximum Simulated Likelihood (MSL) methods are used while in the Bayesian realm, Markov Chain Monte Carlo (MCMC) methods are employed. However, one disadvantage of these methods is that model improvement is not sought through incorporating heterogeneity within observed utility component (systematic heterogeneity).

A *third* approach to accommodate heterogeneity is to undertake an endogenous segmentation or develop finite mixture model. It was introduced by Kamakura and Russel (1989)

and after its introduction, the model and its different variants are applied to varied empirical contexts including mode choice (Bhat, 1997), vehicle ownership (Anowar et al., 2014), lifestyle preferences (Walker and Li, 2007), air carrier choice (Drabas and Wu, 2013; Wen and Lai, 2010), and residence and workplace location (Waddell et al., 2007). Rather than allocating observations exclusively, the probability of belonging to different segments is computed, and segment-specific choice models are estimated. The model estimation process begins with two segments and the number is increased one segment at a time until no improvement in data fit can be obtained with the additional segment. In this framework, the segment membership is a function of multivariate set of exogenous variables (as opposed to a select subset) within a closed form approach. The model estimation can be undertaken using Full Information Maximum Likelihood (FIML) or the Expectation Maximization (EM) approach (Sobhani et al., 2013). The approach is elegant and has received increasing attention in recent years (see Wafa et al., 2015; Xie et al., 2012; Eluru et al., 2012a; Tang and Mokhtarian, 2009).

Finally, a *fourth* approach addresses the homogeneity assumption by formulating joint/multivariate modeling frameworks. To enhance our understanding of the dependent variable of interest, in this approach, we draw additional information for an observation (usually an individual or household) by augmenting with another dependent variable. The approach is well recognized in the transportation literature for its application to residential self-selection. Specifically, choice dimensions (such as mode choice or vehicle fleet size) are studied in conjunction with residential location choice. The emphasis of the approach is on accounting for unobserved factors that affect these dimensions simultaneously allowing us to parse the influence of exogenous variables accurately. In fact, the approach can be visualized as an enhancement of the first approach discussed. In addition to the mutually exclusive sampling, the approach considers the choice mechanism for the clustering variable and couples it with the actual dependent variable of interest. For example, in the case of modeling travel mode choice or vehicle ownership, the decision is coupled with residential location choice (Pinjari et al., 2007; Bhat and Guo, 2007; Pinjari et al., 2011; Paleti et al., 2013) and examined as a joint residential location and mode choice/vehicle ownership. The well recognized switching regime model (see Bhat and Eluru, 2009) also falls within this realm.

## 1.2 Current Study Contributions

While there have been multiple research efforts comparing and contrasting the first three approaches (Greene and Hensher, 2013; Shen, 2009; Greene and Hensher, 2003; Magidson and Vermunt, 2002), there is no effort to compare the fourth approach with the first three approaches. The rationale behind the comparison effort is to evaluate if adding additional complexity to the model system with additional dependent variables provides adequate improvement in data fit to warrant the additional dimensions and complexity that arises. To be sure, the comparison is not straightforward. Within the fourth approach, the estimated measures of fit are based on the joint distribution of dependent variables considered while in the other three approaches, the analysis is based on a single dependent variable. Hence, a detailed post-processing effort of is necessary to

generate comparable measures of fit across these approaches. In our study, we clearly elaborate on these measures and provide a comparison across the four approaches by computing a host of comparison metrics (since coefficients of the estimated models cannot be compared). The proposed comparison is undertaken for household vehicle fleet size decisions for Greater Montreal Area (GMA) region using a comprehensive set of explanatory variables.

The remainder of the paper is organized as follows. Section 2 provides details of the econometric model frameworks used in the analysis. In Section 3, description of the data source and sample formation procedures are presented. Model Estimation results are discussed in Section 4. In Section 5, comparison results for both estimation and validation samples are provided. The policy evaluation results are also presented in Section 5. Finally, Section 6 concludes the paper.

## 2. ECONOMETRIC FRAMEWORK

In this section, we briefly provide the details of the econometric framework of the models considered for examining household vehicle ownership levels in our study. We introduce the traditional ordered logit (OL) model, then discuss the mixed ordered logit (MOL) model (approach 2), latent segmentation based ordered logit (LSOL) model (approach 3), and copula based joint residential location (MNL) and vehicle ownership model (OL) (approach 4).

### 2.1 Ordered Logit (OL) Model

Let  $h$  ( $h = 1, 2, \dots, H$ ) and  $j$  ( $j = 1, 2, \dots, J$ ) be the indices to represent decision makers (households) and vehicle fleet sizes, respectively. In the traditional OL model, vehicle ownership levels ( $y_h$ ) are assumed to be associated with an underlying continuous latent variable ( $y_h^*$ ). This latent variable is typically specified as the following linear function:

$$y_h^* = \alpha z_h + \varepsilon_h, \quad y_h = j, \quad \text{if } \tau_{j-1} < y_h^* < \tau_j \quad (1)$$

where,  $y_h^*$  is the latent propensity for household  $h$  choosing a vehicle ownership level  $j$ ,  $z_h$  is a vector of exogenous variables,  $\alpha$  is a vector of coefficients to be estimated and  $\varepsilon_h$  is a random disturbance term assumed to be standard logistic. The latent propensity  $y_h^*$  is mapped to the observed ownership levels  $y_h$  by  $\tau$  thresholds ( $\tau_0 = -\infty, \tau_J = +\infty$ ) with the following ordering conditions:  $(-\infty < \tau_1 < \tau_2 < \dots < \tau_{J-1} < +\infty)$ . Given these relationships across the different parameters, the resulting probability expression takes the following form:

$$P_{hj}(y_h = j) = \Lambda(\tau_j - \alpha z_h) - \Lambda(\tau_{j-1} - \alpha z_h) \quad (2)$$

where,  $\Lambda(\cdot)$  is the standard logistic cumulative distribution function (see Train, 2003; Greene and Hensher, 2010 for more details).

### 2.2 Mixed Ordered Logit (MOL) Model

Mixed OL model accommodates unobserved heterogeneity in the effect of exogenous variables on household vehicle ownership levels in the latent vehicle owning propensity function (see Train, 2003; Greene and Hensher, 2010 for more details):

$$y_h^* = (\alpha + \gamma)z_h + \varepsilon_h \quad (3)$$

$\gamma$  is a vector of unobserved factors moderating the influence of attributes in  $z_h$  on the vehicle owning propensity for household  $h$ . In the current paper, we assume that the  $\gamma$  elements are independent realizations from normal distribution.

### 2.3 Latent Segmentation based Ordered Logit (LSOL)

Let us consider  $S$  homogenous segments of households (the optimal number of  $S$  is to be determined). In this approach, households are probabilistically assigned to the segments for the segmentation model. The utility for assigning a household  $h$  ( $1, 2, \dots, \dots, H$ ) to segment  $s$  is defined as:

$$U_{hs} = \beta_s z_h + \xi_{hs} \quad (4)$$

$z_h$  is a column vector of attributes that influences a household's propensity of belonging to segment  $s$ ,  $\beta_s$  is a corresponding column vector of coefficients, and  $\xi_{hs}$  is an idiosyncratic random error term assumed to be identically and independently Type 1 Extreme Value distributed across households  $h$  and segment  $s$ . Then the probability that household  $h$  belongs to segment  $s$  is given as:

$$P_{hs} = \frac{\exp(\beta_s z_h)}{\sum_s \exp(\beta_s z_h)} \quad (5)$$

Within the latent segmentation approach, the unconditional probability,  $P_h(j)$  of household  $h$  choosing auto ownership level  $j$  is given as:

$$P_h(j) = \sum_{s=1}^S (P_h(j)|s)(P_{hs}) \quad (6)$$

The log-likelihood (LL) function for the entire dataset with appropriate conditional probability,  $P_h(j)|s$  (in our case,  $P_h(j)|s = \Lambda(\psi_{s_k} - \alpha_s x_h) - \Lambda(\psi_{s_{k-1}} - \alpha_s x_h)$  where  $\psi$  represents thresholds) for ordered regime is provided below:

$$L = \sum_{h=1}^H \log(P_h(j)) \quad (7)$$

### 2.4 Copula based Joint MNL-OL Model

In our empirical analysis, in addition to vehicle ownership, we considered another dependent variable – residential location choice. The former is modeled using OL structure, and the latter is modeled using MNL structure. These two dependent variables are jointly analyzed using a copula approach (see Yasmin et al., 2014; Rana et al., 2010; Portoghesi et al, 2011 for a similar modeling technique in a different context). The copula approach is gaining wide applicability in recent years

amongst travel behavior researchers (Rashidi and Mohammadian, 2016; Ermagun et al., 2015; Zou et al., 2014).

#### 2.4.1 Residential Location Model Component

Let  $r$  ( $r = 1, 2, \dots, R$ ) be the index representing residential location choice of households  $h$  ( $1, 2, \dots, H$ ). Following random utility theory, the propensity of a household  $h$  choosing a residential location  $r$  takes the form of:

$$u_{hr}^* = \beta_r x_{hr} + \xi_{hr} \quad (8)$$

where,  $x_{hr}$  is a vector of exogenous variables,  $\beta_r$  is a vector of unknown parameters specific to residential location  $r$  and  $\xi_{hr}$  is an idiosyncratic error term (assumed to be standard type-I extreme value distributed) capturing the effects of unobserved factors on the propensity associated with residential location  $r$ . A household  $h$  is assumed to choose location type  $r$  if and only if the following condition holds:

$$u_{hr}^* > \max_{l=1,2,\dots,r, l \neq r} u_{hl}^* \quad (9)$$

The condition presented in Equation (9) can be equivalently represented as a series of binary outcome models for each location choice,  $r$  (see Lee, 1983). For example, let  $\eta_{hr}$  be a dichotomous variable with  $\eta_{hr} = 1$  if a household  $h$  ends up choosing residential location  $r$  and  $\eta_{hr} = 0$  otherwise. Now, let us define  $v_{hr}$  as follows:

$$v_{hr} = \xi_{hr} - \left\{ \max_{l=1,2,\dots,r, l \neq r} u_{hl}^* \right\} \quad (10)$$

By substituting the right side for  $u_{hr}^*$  from Equation (8) in Equation (9), we can write:

$$\eta_{hr} = 1 \text{ if } \beta_r x_{hr} + v_{hr} > 0 \quad (11)$$

An assumption of independent and identical Type 1 Gumbel distribution for  $\xi_{hr}$  results in a logistic distributed  $v_{hr}$ . Consequently, the probability expression for the corresponding residential location model resembles the multinomial logit probability expression:

$$\Lambda_r(\beta_r x_{hr}) = Pr(v_{hr} > -\beta_r x_{hr}) = \frac{\sum_{l \neq r} \exp(\beta_r x_{hl})}{\exp(\beta_r x_{hr}) + \sum_{l \neq r} \exp(\beta_r x_{hl})} \quad (12)$$

#### 2.4.2 Vehicle Ownership Model Component

Considering the vehicle ownership levels to be ordered, the probability expression for household  $h$  owning vehicle fleet size  $j$  in a residential location  $r$  takes the following form:

$$Pr(y_{hr} = j_r) = \Lambda_r(\tau_{r,j} - \alpha_r z_{hr}) - \Lambda_r(\tau_{r,j-1} - \alpha_r z_{hr}) \quad (13)$$

where,  $\Lambda_r(\cdot)$  is the standard logistic cumulative distribution function.

### 2.4.3 The Joint Model

The vehicle ownership and residential location component discussed in the previous two subsections may be brought together in the following equation system:

$$\begin{aligned} \eta_{hr} &= 1 \text{ if } \beta_r x_{hr} > v_{hr} \\ y_{hr}^* &= \alpha_r z_{hr} + \varepsilon_{hr}, \quad y_{hr} = 1[\eta_{hr} = 1]y_{hr}^* \end{aligned} \quad (14)$$

In constructing the copula dependency, the random variables ( $v_{hr}$  and  $\varepsilon_{hr}$ ) are transformed into uniform distributions by using their inverse cumulative distribution functions, which are then coupled or linked as a multivariate joint distribution function by applying the copula structure. Let us assume that  $\Lambda_{vr}(\cdot)$  and  $\Lambda_{\varepsilon r}(\cdot)$  are the marginal distribution of  $v_{hr}$  and  $\varepsilon_{hr}$ , respectively and  $\Lambda_{vr,\varepsilon r}(\cdot, \cdot)$  is the joint distribution of  $v_{hr}$  and  $\varepsilon_{hr}$ . Subsequently, a bivariate distribution  $\Lambda_{vr,\varepsilon r}(v, \varepsilon)$  can be generated as a joint cumulative probability distribution of uniform [0,1] marginal variables  $U_1$  and  $U_2$  as below:

$$\begin{aligned} \Lambda_{vr,\varepsilon r}(v, \varepsilon) &= Pr(v_{hr} < v, \varepsilon_{hr} < \varepsilon) \\ &= [\Lambda_{vr}^{-1}(U_1) < v, \Lambda_{\varepsilon r}^{-1}(U_2) < \varepsilon] \\ &= [U_1 < \Lambda_{vr}(v), U_2 < \Lambda_{\varepsilon r}(\varepsilon)] \end{aligned} \quad (15)$$

The joint distribution (of uniform marginal variable) in Equation (15) can be generated by a function  $C_{\theta h}(\cdot, \cdot)$  (Sklar, 1973), such that:

$$\Lambda_{vr,\varepsilon r}(v, \varepsilon) = C_{\theta h}(U_1 = \Lambda_{vr}(v), U_2 = \Lambda_{\varepsilon r}(\varepsilon)) \quad (16)$$

where  $C_{\theta h}(\cdot, \cdot)$  is a copula function and  $\theta_h$  the dependence parameter defining the link between  $v_{hr}$  and  $\varepsilon_{hr}$ .

### 2.4.4 Estimation Procedure

The joint probability that the household  $h$  residing in a residential location type  $r$  and owning a fleet size of  $j$ , from Equation 9 and 14, can be written as:

$$\begin{aligned} &Pr(\eta_{hr} = 1, y_{hr} = j_r) \\ &= Pr\left\{(\beta_r x_{hr} > v_{hr}), \left((\tau_{r,j-1} - \alpha_r z_{hr}) < \varepsilon_{hr} < (\tau_{r,j} - \alpha_r z_{hr})\right)\right\} \\ &= Pr\left((\beta_r x_{hr} > v_{hr}), (\varepsilon_{hr} < \tau_{r,j} - \alpha_r z_{hr})\right) \\ &\quad - Pr\left((\beta_r x_{hr} > v_{hr}), (\varepsilon_{hr} < \tau_{r,j-1} - \alpha_r z_{hr})\right) \\ &= \Lambda_{\varepsilon r}(\tau_{r,j} - \alpha_r z_{hr}) - \Lambda_{\varepsilon r}(\tau_{r,j-1} - \alpha_r z_{hr}) - \left(Pr[v_{hr} < -\beta_r x_{hr}, \varepsilon_{hr} < \right. \\ &\quad \left. (\tau_{r,j} - \alpha_r z_{hr})\right] - Pr[v_{hr} < -\beta_r x_{hr}, \varepsilon_{hr} < (\tau_{r,j-1} - \alpha_r z_{hr})\right]) \end{aligned} \quad (17)$$

The joint probability of Equation (17) can be expressed by using the copula function as:

$$\begin{aligned} &Pr(\eta_{hr} = 1, y_{hr} = j_r) \\ &= \Lambda_{\varepsilon r}(\tau_{r,j} - \alpha_r z_{hr}) - \Lambda_{\varepsilon r}(\tau_{r,j-1} - \alpha_r z_{hr}) \\ &\quad - [C_{\theta h}(U_{h,j}^r, U_h^r) - C_{\theta h}(U_{h,j-1}^r, U_h^r)] \end{aligned} \quad (18)$$



where  $U_{h,j}^r = \Lambda_{er}(\tau_{r,j} - \alpha_r z_{hr})$ ,  $U_h^r = \Lambda_{vr}(-\beta_r x_{hr})$ .

Thus the likelihood function with the joint probability expression in Equation (18) for vehicle ownership and residential location outcomes can be expressed as:

$$L = \prod_{h=1}^H \left[ \prod_{r=1}^R \prod_{j=1}^J \{Pr(\eta_{hr} = 1, y_{hr} = j)\}^{\omega_{hrj}} \right] \quad (19)$$

where,  $\omega_{hrj}$  is dummy with  $\omega_{hrj} = 1$  if the household  $h$  resides in residential location  $r$  and own vehicle fleet size of  $j$  and 0 otherwise. All the parameters in the model are then consistently estimated by maximizing the logarithmic function of  $L$ . The parameters to be estimated in the model are:  $\beta_r$  in the MNL component,  $\alpha_r$  and  $\tau_{r,j}$  in OL component. In our analysis, we employ six different copulas structure - the Gaussian copula, the Farlie-Gumbel-Morgenstern (FGM) copula, and a set of Archimedean copulas including Frank, Clayton, Joe and Gumbel copulas (a detailed discussion of these copulas is available in Bhat and Eluru, 2009). Once the parameters are obtained, the equation for calculating the probability (as marginal of the joint distribution) for the  $j^{th}$  alternative from the joint model is as follows:

$$P(j) = \sum_{\forall r} Pr(\eta_{hr} = r, y_{hr} = j) \quad (20)$$

In our case, equation (20) is a summation of as many terms as the number of residential location alternatives for each vehicle fleet size ( $j$ ). Please note that restricting the copula structure to have no correlation between the error terms of residential location and vehicle ownership choices would result in a residential location cluster based ordered logit or exogenous segmentation based ordered logit (ESOL) model. All the parameters in all the models are consistently estimated by maximizing the log-likelihood function, which is accomplished using the GAUSS matrix programming language.

### 3. CASE STUDY

#### 3.1 Endogeneity in Vehicle Ownership

Private automobile ownership (fleet size and composition) and residential location choice are two household decisions that have significant impact on transportation outcomes. The interrelationship between these two decision processes is well established in the extant travel behavior literature (de Abreu e Silva et al., 2012a; Rashidi et al., 2011; van Acker and Witlox, 2010) and has received considerable attention from the transportation community over the years. One stream of studies assumes one-way causal relationship between vehicle ownership and residential location and treats land use characteristics as purely exogenous factors in models of vehicle ownership (see Anowar et al., 2016; Kowald et al., 2016; Anowar et al., 2014; Scott and Axhausen, 2006). The other stream of studies argues that considering residential location as merely an exogenous variable may provide erroneous indications of the true impacts of land use on vehicle ownership. This might be

due to a phenomenon referred to as self-selection or residential sorting where households with a proclivity towards certain lifestyle (for example, desire or lack of desire to own vehicles), choose to “self-select” or reside in a location conducive to their preferred lifestyle and travel inclinations (de Vos and Witlox, 2016; Chatman, 2014; Chen and Lin, 2011). In other words, vehicle ownership levels of households might be endogenous to their choice of location of residence. In econometric theory, the endogeneity bias occurs from the common unobserved factors affecting the two choice processes. To accommodate for the interdependency, researchers have jointly modeled residential location choice and vehicle ownership levels. Bhat and Guo (2007) and Pinjari et al. (2008) both used multidimensional models to capture endogeneity between vehicle ownership and other decision processes.

A list of studies investigating endogeneity between vehicle ownership and residential location choice is presented in Table 1(a) and 1(b). Table 1(a) includes studies that explicitly consider residential location as a decision variable to address the endogeneity issue while Table 1(b) contains studies that examine endogeneity by considering the interaction of built environment and vehicle ownership. Both tables provide information on the study area, methodology employed, measures of endogenous travel behavior investigated, and the exogenous variable categories considered in the analysis. The only difference between the tables is in terms of explanatory variables; policy variable category is included in Table 1(a) while Table 1(b) includes attitude/lifestyle variable. The following observations may be made from the tables. *First*, most of the evidence comes from North America, particularly USA. Very few studies are in the Asian, European, and South American contexts. *Second*, majority of the studies are cross-sectional in nature, with Cao et al. (2007) and Aditjandra et al. (2012) being the only two exceptions (quasi-longitudinal). *Third*, for addressing the self-selection issue, the two most prevalent methodologies employed in the studies are the multidimensional models and the Structural Equation Models (SEM). *Fourth*, examination of these studies further revealed that residential location choice options incorporated in the models are of two types. The choices are formulated by either considering the smallest available geographic unit, such as the traffic analysis zones (TAZ) or the census tract (CT) (Bhat and Guo, 2007; Salon, 2009) or aggregating the city neighborhoods, based on selected attributes, into a small number of interpretable geographic units (Bhat et al., 2013; Guerra, 2015). For example, Bhat et al. (2013) and Bhat et al. (2014) used density of the census blocks in which the household resides to create seven location choice alternatives while Bhat and Eluru (2009) used factor and cluster analysis to classify TAZs into neo-urbanist and conventional neighborhoods.

### 3.2 Data Source and Preparation

Montreal is the second largest Census Metropolitan Area (CMA) in Canada characterized by a diverse urban form and a unique heterogeneous multimodal transportation system comprised of metro, commuter train, and an extensive bus service. The city has a relatively high share of transit ridership (for a North American city) (Eluru et al., 2012b). With 500 km of recreational and on-street bicycle paths, Montreal is also one of North America’s top destinations for urban cycling.

The urban region size with its land use mix, fairly well-developed public transportation system, and active transportation infrastructure and culture makes Montreal an ideal subject to investigate the issue of residential self-selection in the context of household vehicle ownership levels.

The primary data source used in the current analysis is the 2008 cross-sectional Origin-Destination (O-D) survey of Greater Montreal Area (GMA). The O-D surveys are conducted every five years and they are the primary source of information on individual mobility patterns in the GMA region. Several socioeconomic characteristics of both individuals and households are recorded, including age, gender, work status, license status, and number of household members. The O-D data was augmented with a host of secondary GIS data sources and Census data. The sample employed in the empirical analysis was prepared in several steps. From the 66,124 records, a random sample was drawn which comprised approximately 16,000 households, of which 10,214 data records were used for estimation and 5,455 data records were set aside for model validation. The sampling exercise was undertaken primarily to reduce the data compilation burden using ArcGIS. Additionally, considering large data samples for model estimation could also result in inflated parameter significance. The random sampling process was carefully undertaken to ensure that estimated sample dependent variable distribution matched with the full sample dependent variable distribution.

### **3.3 Residential Cluster Generation**

GMA is comprised of 878 census tracts (CT). Based on a thorough review of previous literature, a comprehensive list of urban form and land use variables were generated using ArcGIS platform for each of these tracts. Then, residential location alternatives for the households were created by clustering these CTs using *k*-means clustering technique since there was no existing urban/suburban/rural typology provided in the survey data (similar technique was used in Salon, 2015; Manaugh et al., 2010; Pinjari et al., 2008). *k*-means cluster analysis groups each tract into one of a pre-determined number of clusters based on selected variables such that internal similarity is maximized while similarities between groups are minimized. The variables used to define clusters are: (1) population density (population per acre), (2) job density (number of jobs per 15-65 year aged population), (3) number of detached households in CT, (4) number of transit, bike, and walk commuters, and (5) number of dwelling units built before 1946. The selection of the clustering variables was guided by their significance in similar previous research in creating neighborhood typologies (Salon, 2015; Patterson et al., 2014; Miranda-Moreno et al., 2012; Harding et al., 2014). Population density is a good indicator of urbanity and land use mix while job density captures labor demand (Hastings, 2003). Share of single-detached dwellings in a tract is a good marker of homogeneity/heterogeneity of land use developments. The share of active and transit mode users is included as a proxy for accessibility to transit. Finally, the share of dwellings built before 1946 provides an indication of historical core of the city (de Vos and Witlox, 2016; Patterson et al., 2014). Moreover, researchers have reported that percentage of buildings built before Second World War is likely to be correlated with urban form variables (Boarnet and Sarmiento, 1998; Vance and Hedel, 2007).

The clustering procedure was carried out in SPSS. After classifying the tracts, using ArcGIS, visual inspection of the clustering result was conducted as a “sanity check” for obvious misclassification or dubious classification, which might simply be due to inadequate number of clusters specified (Lin and Long, 2008). In our case, the optimal number of cluster categories extracted was found to be 4 (see Figure 1). Then households were assigned to these clusters based on which CT they fall in geographically. After reviewing these clusters (local knowledge of the region was very useful in this regard), each cluster was labelled to represent its characteristics. A brief description of each of the final clusters is provided below:

- Cluster 1: Urban Core – represents mostly the downtown core and central neighborhoods, with the highest values in each of the five input variables. This includes the historic core city, and a very heterogeneous land use mix.
- Cluster 2: Inner Suburb – this is the intermediate residential location type with all values of the clustering variables being equal or very close to the average.
- Cluster 3: Outer Suburb – where all attributes have values slightly below the average.
- Cluster 4: Suburban/Rural – characterized by the lowest values in all categories, which is also referred to as the periphery.

Table 2 lists the characteristics that were used to classify census tracts together with their average values in each resulting neighborhood type.

### 3.4 Independent Variables and Descriptive Statistics

In the current study, a comprehensive set of exogenous attributes were considered to examine vehicle ownership levels and residential location choices of households. The independent variables can be broadly classified into three categories: (1) household socio-demographic characteristics, (2) land use and built environment characteristics, and (3) transit accessibility measures. To account for the impact of worker’s transit accessibility at work locations on household’s vehicle fleet size decision, we created the interaction of the number of worker variable with varying degrees of transit accessibility (no access, low, medium, and high access). In addition to the regional land use characteristics, the local attributes in the vicinity of the location of the household were also compiled for our analysis. This was achieved by creating 600m circular buffer<sup>1</sup> around household residential location. The list of the variables and their definitions are presented in Table 3(a).

Car ownership levels were classified as no car, one car, two cars, and three or more cars. The distribution of auto ownership levels in the estimation sample were as follows: 20.6% of the households were carless, 43.1% owned one car, 29.1% owned two cars, and 7.2% of the households had a fleet size in excess of two cars. Distribution of households in terms of residential location were as follows: 26.8% resided in the urban core, 25.1% in the inner ring area, 28.8% in the outer ring area, and 19.3% in the suburban area. Moreover, the auto ownership descriptive

---

<sup>1</sup> Buffers were established around household geocoded locations with 600m radius. In the earlier literature, the acceptable walking distance to transit stops and stations is often assumed to be 400m (Larsen et al., 2010). We employed a larger buffer than 400m to allow for the low-density developments in Canadian cities that might require people to walk further to reach transit stations from their households.

analysis indicated an average ownership of 1.27 vehicles per household. Vehicle ownership distribution across residential location clusters are presented in Figure 2. From the figure, it can be observed that households with larger fleet sizes live predominantly in the suburbs while households in the central areas have fewer cars. Some other salient characteristics of the sample are: the majority of the households (60.8%) reside in medium income CTs, two-thirds have at least one male adult (67%), one full-time employed member (64.6%), about three-quarters have at least one part-time worker, nearly one-third of the households have at least one child and one-third of the households have at least one retiree. We can also observe from Table 3(b) that the proportions of different variables vary substantially across different outcomes of the residential location and auto ownership decisions. Note that the percentages sum to 100% for each exogenous variable across the vehicle ownership columns and across residential location clusters columns.

## 4. EMPIRICAL ANALYSIS

### 4.1 Model Specification and Overall Measures of Fit

The empirical analysis involved a series of model estimations. To analyze household vehicle ownership, we estimated five models: (1) traditional ordered logit (OL) model, (2) exogenous segmentation residential location cluster based ordered logit model (ESOL) – 4 cluster specific OL models, (3) mixed ordered logit (MOL) model, (4) latent segmentation based ordered logit model with two (LSOLII), and three (LSOLIII) segments, (5) a copula based joint residential location and vehicle ownership model. The estimation process is relatively straightforward for models (1), (2) and (3). The log-likelihood (parameters) for these models are as follows: OL [–9101.60 (27)], ESOL [–9075.30 (67)], and (3) MOL [–9096.12 (28)]. Estimation of models (4) and (5) involve multiple steps. For the latent segmentation modeling approach, the model estimation process began with a model considering two segments. The final number of segments was determined by adding one segment at a time until further addition did not enhance intuitive interpretation and data fit. Finally, the number of segments corresponding to the lowest value of Bayesian Information Criterion (BIC)<sup>2</sup> was considered as the appropriate number of segments. However, it should be noted that the decision regarding the optimal number of classes should be made considering the significance of the number of parameters and the interpretability of the model. The BIC (number of parameters estimated) values for the LSOL model with two and three segments were, 18104 (42) and 18266 (43), respectively. Therefore, we selected two segments as the appropriate number of segments. From here on latent segmentation model refers to the two segment latent class model.

---

<sup>2</sup> The Bayesian Information Criterion (BIC) for a given empirical model is equal to  $[-2(\text{LL}) + K \ln(Q)]$ , where (LL) is the log-likelihood value at convergence, K is the number of parameters, and Q is the number of observations. BIC is found to be the most consistent Information Criterion (IC) for correctly identifying the appropriate number of segments in latent segmentation models (see Nylund et al., 2007). Moreover, several research studies have employed BIC to compare models because the test statistic accommodates data fit and number of parameters in identifying the best model (Burnham and Anderson, 2004; Kuha, 2004).

For the joint copula based MNL-OL model, we estimated models considering six different copula structures: (1) Gaussian, (2) FGM, (3) Clayton, (4) Gumbel, (5) Frank and (6) Joe. Copula models that allow for different dependency structures for different residential location choice and vehicle ownership level combinations were also estimated. The lowest BIC value was obtained for a combination model of Joe-FGM copulas. The log-likelihood value at convergence for the Joe-FGM model structure was found to be  $-21677.79$  (81) while the log-likelihood at convergence for the independent model structure that ignores any potential copula dependency was  $-22052.23$  (99). The BIC values for the Joe-FGM model and the independent model were  $44103$  and  $45018$ , respectively. The results support the presence of common unobserved factors influencing location and vehicle ownership choice processes. In terms of dependency, note that a positive parameter indicates that the dependencies caused by the common unobserved factors for the specific location clusters are positive, and a negative parameter indicates that the dependencies are negative. In our model, Joe copula dependency structure is associated with urban core and inner ring neighborhoods while FGM dependency structure is associated with outer ring and suburban locations. Joe copula structure can only account for positive dependence and offers a stronger right tail dependence. The magnitude of dependence decreases from the urban core location to the inner ring location. FGM can accommodate both positive and negative values. The parameter for FGM copula was found to be at the negative limit implying that weak dependency was captured using this copula. The result suggests that a household that makes a choice to reside in urban core is also likely to own more vehicles. The result might seem counterintuitive at first glance. However, it might be attributed to, among other things, unobserved factors characterizing households who enjoy the amenities and activities that a city core like Montreal has to offer but at the same time, enjoy the social status of owning a large vehicle fleet despite living in an area with high transit accessibility. Similar results were reported by Li et al. (2010), Cullinane and Cullinane (2003), Innocenti et al. (2013), Sanko et al. (2014), and Bhat and Eluru (2009).

We can compare the four models of vehicle ownership that are non-nested using the BIC values. The BIC values for each of the four models are: OL [18452], ESOL [18409], MOL [18450], and LSOLII [18104]. From these comparisons, clearly the latent segmentation framework outperforms all other approaches. The BIC value for joint copula model is not comparable to the other models because it has two dependent variables in the computation of LL and BIC. Hence, based on the final set of convergence estimates, using Equation (20), we generated the marginal for vehicle ownership dimension based on the joint probability prediction. The LL generated was  $-8481.99$  and the corresponding BIC value was  $17472$ . Clearly, the numbers indicate that the vehicle ownership component estimated from the joint copula based model outperforms all other frameworks.

## 4.2 Model Results

We present the results of MOL model in Table 4, ESOL model in Table 5, LSOLII model in Table 6, and copula based model in Table 7(a) (residential cluster choice) and Table 7(b) (vehicle ownership). Note that we are estimating four different model systems and each system is

comprised of different components. For example, in the MOL model, we estimate the household level disturbance of the mean effects. The LSOLII model has two parts – the segmentation component where households are assigned to different segments based on exogenous attributes and segment specific vehicle ownership component. In the copula based model, we have residential cluster choice and the corresponding vehicle ownership components. In addition, we also estimate dependence effects between the two choice components. For the sake of brevity, we provide only brief explanation of the different components of all the models, and discuss the vehicle ownership component of all of these models together. The model estimation process began with the same explanatory variables data pool and the final specification was based on a systematic process of removing statistically insignificant variables at 95% confidence level and combining variables when their effects were not significantly different.

#### *4.2.1 Segmentation Component of Latent Segmentation Based Ordered Logit (LSOLII) Model*

From the segment shares, it is observed that the likelihood of households belonging to segment-2 is the highest (68%). Further, the car ownership probabilities for households, conditional on their belonging to a particular segment, indicate that the two segments exhibit very distinct car ownership profiles. For example, the households belonging to segment 1 are less likely to own zero cars (only 12%), while households assigned to segment 2 are less likely to own 3 or more cars (only 8%). The probability that a household belongs to either one of these two segments is found to be influenced by land use variables including area of the census tract where the household is located, land use mix, number of detached, rented households in the CT, and urban core residential cluster.

#### *4.2.2 Residential Cluster Choice Component of Copula based MNL-OL Model*

In presenting the effects of the exogenous variables in the MNL-OL joint model specification, we will restrict ourselves to the discussion of the Joe-FGM specification. We found that households with children are more likely to locate themselves in lesser density neighborhoods. Similar residential location choice preference was also observed for households with higher number of retirees and driving license holders. Moreover, households comprised of workers with reduced transit accessibility at work, are attracted to peripheral location (outer ring and suburban) while workers with high transit accessibility prefer high density areas. On the other hand, households with higher number of male adults or young adults or students or part-timers are more likely to choose high density areas for residing. These results suggest that households composed of members from these demographic groups might be interested in urban lifestyles and are more environmentally cautious.

#### *4.2.3 Vehicle Ownership Component (All Models)*

All the variables in the vehicle ownership component have similar signs in MOL, ESOL, LSOLII, and Copula OL models. The same factors were often found to influence the vehicle ownership decision across all models. For example, larger households (those with higher number of male

adults or with adults between the age of 19-64 years) have a higher propensity to own more vehicles, presumably to reduce the vehicle resource constraints on members, particularly if the household is located outside of urban core (similar results were reported in Beige and Axhausen, 2008). Irrespective of location, households with higher number of full- and part-time workers are more likely to own more cars. Some variables were only significant in one of the segments. For instance, presence of teenaged children (15-19 years of age), number of female adults, and number of middle-aged adults had a positive effect on vehicle owning propensity in Segment-1. At 18, teenagers are allowed to drive alone and thus, households might acquire extra cars to allow them to drive independently (Prillwitz et al., 2006). Moreover, the impact of some of the exogenous variables varied both in sign and magnitude between the two segments highlighting the presence of population heterogeneity. For instance, we observed that households in Segment-1 have a higher likelihood of owning less cars with an increase in the number of retirees in the household while the opposite effect is observed for Segment-2. In the MOL and copula based models, number of retirees had similar effects. Households living in urban core have a higher likelihood of owning more vehicles if toddlers (0-4 years) are present. Households might enjoy the extra flexibility that personal automobiles offer in terms of traveling with children (for example, dropping children off to day-care, school and/or participate in wide variety of leisure activities) and hence, are more inclined towards owning more vehicles (Nolan, 2010), even when located in dense urban areas.

All of the land use measures negatively impacted vehicle ownership except number of driver commuters, number of detached households, and median income level of census tract. There are two plausible explanations for the impact of number of driver commuters in CT: these CTs have low population density and are not well served by transit; also, the accessibility at the job locations via non-auto mode for these commuters are poor, thereby increasing the likelihood of owning more cars (Chen et al., 2008; Salon, 2015). This is further corroborated by our own finding that households with workers who have little or no transit accessibility at their place of work tend to own multiple vehicles. For households located in high income CT, the propensity to own multiple vehicle is normally distributed with a mean of 0.463 and standard deviation of 0.077<sup>3</sup>. In addition to these variables, the results from the copula model indicated that increased population, job, and dwelling density have a negative impact on vehicle fleet size decision of households, particularly when they are located in the outlying areas. Our finding is in line with the results reported in the extant travel behavior literature - increased density helps reduce vehicle ownership levels (see Chen et al., 2008; Li et al., 2010; Schimek, 1996; Dargay and Hanly, 2007).

The transit attributes found significant was number of bus stops, bus destination diversity, and number of commuter rail stops within the household buffer. As expected, the effect was negative indicating that better transit service obviates the need to have large fleet size, more so if the households are located outside of urban core area.

---

<sup>3</sup>The estimates indicate the distribution is primarily positive as can be observed by the larger mean relative the standard deviation.



## 5. MODEL COMPARISON

We evaluated the performance of the estimated models (MOL, ESOL, LSOLII (obtained from Step 1), and copula OL model (obtained from Step 2) using both aggregate and disaggregate measures of fit. At the aggregate level, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are computed by comparing the predicted and observed shares of vehicle ownership levels. The reader would note that aggregate level statistics are based on just 4 values (aggregated predictions of the 4 vehicle ownership categories considered – 0 car, 1 car, 2 cars, and 3 or more cars) and thus are useful to identify large scale errors only. At the disaggregate level, we used the log-likelihood at convergence (for estimation sample), predictive log-likelihood (for validation sample), and BIC. We compute these measures for the full as well as specific subsamples within the estimation and validation datasets. The subsamples were created based on the following variables: median income of census tract (medium), presence/absence of full-time worker, residential location cluster (urban core), and presence/absence of bus stops. Please note that similar to the estimation dataset, the validation was extracted using random sampling procedure from the base O-D survey data and was set aside during model estimation.

### 5.1 Estimation Sample

The validation results for the estimation sample are presented in Table 8. We can observe that the copula OL (Joe-FGM) model performs reasonably well at the disaggregate level for the entire sample (LL = -8481.99) and for specific subsamples compared to LSOLII, MOL, and ESOL models. The performance essentially provides support to the notion that adding additional dependent variables to predict the dependent variable of interest provides additional data fit. In our empirical context, the ordering of performance across model frameworks is as follows: copula based OL model, LSOLII model, ESOL model, and MOL model. On the other hand, the comparison of the models at the aggregate level is far from conclusive. We found that LSOLII and MOL models perform better in terms of RMSE and MAPE (only marginally). It is important to recognize that the validation at the disaggregate level is more critical to model comparison relative to the validation at the aggregate level.

### 5.2 Validation Sample

A validation experiment was also carried out in order to ensure that the statistical results obtained above are not a manifestation of over fitting to data. To undertake the validation exercise, we employ the final parameters of the models to predict alternative probabilities for the households in the hold-out sample. For testing the predictive performance of the models, 50 data samples, of about 2000 records each, are randomly sampled from the hold out validation sample consisting of 5,455 records. We evaluate both the aggregate and disaggregate measure of predicted fit by using these 50 different validation samples. For these samples, we present the average measures from the comparison, and also the confidence interval (C.I.)<sup>4</sup> of the fit measures at 95% confidence level. The results for the validation sample are presented in Table 9. The bands computed show

---

<sup>4</sup> Confidence Interval (C.I.) = mean  $\pm$  1.96  $\times$  standard deviation /  $\sqrt{(50)}$

that with an exception of one or two cases, at the disaggregate level, the ordered component of the Joe-FGM copula structure consistently outperformed the independent (ESOL), MOL, and LSOLII models. At the aggregate level, based on the point estimates the Joe-FGM copula did not offer the best fit, however, the confidence bands indicate that there is no statistically significant difference between Joe-FGM copula model and the ESOL model. Hence, there is enough evidence to suggest that copula based MNL-OL model performs significantly better in the empirical analysis compared to the three univariate models estimated in our analysis.

### **5.3 Policy Analysis**

To further evaluate the performance of the alternative model structures, we conducted a policy analysis experiment. More specifically, the percentage changes in vehicle ownership levels were predicted for a unit level increase in the number of full-time workers, part-time workers, 10% increase in bike route length, in pedestrian/bike route length, and in the number of bus stops, within the household buffer. The results are presented in Table 10. The analysis is conducted for the parameter distributions represented by 20 sets of coefficient values to obtain a confidence band at the 95% confidence level (as opposed to only the point estimates).

In general, it is found that the vehicle fleet size changes provided by the copula OL model differ from that of the fleet size changes provided by MOL, ESOL, and LSOL models. The findings are based on the comparison of the 95% confidence level vehicle ownership change distributions (and not just point estimates). While some model prediction differences are small, there are some that are quite substantial (e.g., effect of increase in full-and-part-time workers on 3 or more car ownership levels), suggesting that ignoring dependency across choice dimensions could result in serious over- and/or under-estimation of impacts of changes in exogenous variables. An increase in bus stop numbers within the household buffer decreased multiple vehicle ownership by a small margin. Moreover, our results also suggest that increasing accessibility for both pedestrians and bicyclists by expanding pedestrian and bike routes has a greater impact on reducing vehicle ownership than just increasing bike route length.

## **6. SUMMARY AND CONCLUSIONS**

In this paper, we propose a comparison of different model systems that relax the population homogeneity assumption using vehicle ownership as our case study. To be sure, in this context, several studies have compared and contrasted the single dependent variable model frameworks. However, there is no effort to compare the single dependent approaches with multiple dependent variable approaches. We aim to bridge this gap in the literature by comparing several single dependent variable approaches with a copula based joint modeling approach. The rationale behind the comparison effort is to evaluate if adding additional complexity to the model system with additional dependent variables provides adequate improvement in data fit to warrant the additional dimensions and complexity that arises with it.

Using the 2008 Origin-Destination survey data of Greater Montreal Area (GMA), we estimated and compared five models: (1) traditional ordered logit (OL) model, (2) exogenous

segmentation residential location cluster based ordered logit model (ESOL) – 4 OL models, (3) mixed ordered logit (MOL) model, (4) latent segmentation based ordered logit model with two (LSOLII), and three (LSOLIII) segments, (5) a copula based joint residential location and vehicle ownership model. For the joint model system, in addition to vehicle ownership, we considered another dependent variable – residential cluster location choice which were generated using multivariate *k*-means clustering technique. In the univariate models, the residential location cluster information was used as an exogenous variable. This is a major difference between the univariate models and the multivariate copula model. In light of the difference, several post-processing steps were undertaken to generate comparable model fit measures across these approaches. The performance of the models was examined for set of full sample as well as for specific sub-samples of estimation and validation dataset. In our empirical context, with an exception of one or two cases, both at aggregate and disaggregate levels, the ordered component of the Joe-FGM copula structure consistently outperformed the ESOL, MOL, and LSOLII models. Even in the cases when the univariate models outperformed the copula OL, the difference was found to be marginal. The results indicate that the extra complexity necessary to study the two dependent variables does offer additional fit improvement. However, caution should be exercised in generalizing the study findings since the observation made is for data from one urban region. It would be interesting to test the hypothesis on another city data. Our study could be further enhanced by considering attitudinal factors within the modeling approach for examining residential location choices and decisions about vehicle ownership.

### ACKNOWLEDGMENT

Agence Metropolitaine de Transport (AMT) of Quebec provided us the survey data where a total of 66,124 household records were available. The authors would like to thank both Agence métropolitaine de transport (AMT) and Ministère des Transports du Québec (MTQ) for the data and Professor Ahmed El-Geneidy, Emily Grise, and Genevieve Boisjoly of Transportation Research at McGill (TRAM) for giving access to their GIS data archive. The authors would also like to acknowledge insightful feedback on an earlier version of the manuscript from five anonymous reviewers and the editor.

### REFERENCES

- Aditjandra, P. T., Cao, X. J., & Mulley, C. (2012). Understanding neighbourhood design impact on travel behaviour: An application of structural equations model to a British metropolitan data. *Transportation Research Part A: Policy and Practice*, 46(1), 22-32.
- Anowar, S., Eluru, N., & Miranda-Moreno, L. F. (2016). Analysis of vehicle ownership evolution in Montreal, Canada using pseudo panel analysis. *Transportation*, 43(3), 531-548.
- Anowar, S., Yasmin, S., Eluru, N., & Miranda-Moreno, L. F. (2014). Analyzing car ownership in Quebec City: a comparison of traditional and latent class ordered and unordered models. *Transportation*, 41(5), 1013-1039.

- Bachand-Marleau, J., Larsen, J., & El-Geneidy, A. M. (2011). Much-anticipated marriage of cycling and transit: How will it work? *Transportation Research Record*, 2247, 109-117.
- Beckman, J. D., & Goulias, K. G. (2008). Immigration, residential location, car ownership, and commuting behavior: a multivariate latent class analysis from California. *Transportation*, 35(5), 655-671.
- Beige, S., & Axhausen, K. W. (2007). The ownership of mobility tools during the life course. Paper Presented at the 87<sup>th</sup> Annual Meeting of the Transportation Research Board. Washington, D.C.
- Bhat, C., Paleti, R., Pendyala, R., Lorenzini, K., & Konduri, K. (2013). Accommodating immigration status and self-selection effects in a joint model of household auto ownership and residential location choice. *Transportation Research Record: Journal of the Transportation Research Board*, 2382, 142-150.
- Bhat, C., Eluru, N., & Copperman, R. (2008). Flexible Model Structures for Discrete Choice Analysis. In D. A. Hensher & K. J. Button (Eds.), *Handbook of Transport Modelling (2nd Edition)* (pp. 75-104): Emerald, Inc.
- Bhat, C. R. (1997). An endogenous segmentation mode choice model with an application to intercity travel. *Transportation science*, 31(1), 34-48.
- Bhat, C. R. (2015). A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B: Methodological*, 79, 50-77.
- Bhat, C. R., Astroza, S., Sidharthan, R., Alam, M. J. B., & Khushefati, W. H. (2014). A joint count-continuous model of travel behavior with selection based on a multinomial probit residential density choice model. *Transportation Research Part B: Methodological*, 68, 31-51.
- Bhat, C. R., & Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, 43(7), 749-765.
- 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.
- Boarnet, M. G., & Sarmiento, S. (1998). Can land-use policy really affect travel behaviour? A study of the link between non-work travel and land-use characteristics. *Urban Studies*, 35(7), 1155-1169.
- Brownstone, D., & Fang, H. A. (2014). A vehicle ownership and utilization choice model with endogenous residential density. *The Journal of Transport and Land Use*, 7(2), 135-151.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2), 261-304.
- Cao, J., & Cao, X. (2014). The impacts of LRT, neighbourhood characteristics, and self-selection on auto ownership: evidence from Minneapolis-St. Paul. *Urban Studies*, 51(10), 2068-2087.

- Cao, X., Mokhtarian, P. L., & Handy, S. L. (2007). Do changes in neighborhood characteristics lead to changes in travel behavior? A structural equations modeling approach. *Transportation*, 34(5), 535-556.
- Chamberlain, G. (1980). Analysis of Covariance with Qualitative Data. *The Review of Economic Studies*, 47(1), 225-238.
- Chatman, D. G. (2014). Estimating the effect of land use and transportation planning on travel patterns: Three problems in controlling for residential self-selection. *Journal of Transport and Land Use*, 7(3), 47-56.
- Chen, C., Gong, H., & Paaswell, R. (2008). Role of the built environment on mode choice decisions: additional evidence on the impact of density. *Transportation*, 35(3), 285-299.
- Chen, C., & Lin, H. (2011). Decomposing residential self-selection via a life-course perspective. *Environment and Planning A*, 43(11), 2608-2625.
- Cullinane, S., & Cullinane, K. (2003). Car dependence in a public transport dominated city: evidence from Hong Kong. *Transportation research part D: Transport and environment*, 8(2), 129-138.
- Damant-Sirois, G., Grimsrud, M., & El-Geneidy, A. M. (2014). Whats your type: a multidimensional cyclist typology. *Transportation*, 41(6), 1153-1169.
- Dargay, J., & Hanly, M. (2007). Volatility of car ownership, commuting mode and time in the UK. *Transportation Research Part A: Policy and Practice*, 41(10), 934-948.
- de Abreu e Silva, J., Golob, T., & Goulias, K. (2006). Effects of land use characteristics on residence and employment location and travel behavior of urban adult workers. *Transportation Research Record: Journal of the Transportation Research Board*, 1977, 121-131.
- de Abreu e Silva, J., & Goulias, K. (2009). Structural Equations Model of Land Use Patterns, Location Choice, and Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2135, 106-113.
- de Abreu e Silva, J., Goulias, K., & Dalal, P. (2012a). Structural Equations Model of Land Use Patterns, Location Choice, and Travel Behavior in Southern California. *Transportation Research Record: Journal of the Transportation Research Board*, 2323, 35-45.
- de Abreu e Silva, J., Morency, C., & Goulias, K. G. (2012b). Using structural equations modeling to unravel the influence of land use patterns on travel behavior of workers in Montreal. *Transportation Research Part A: Policy and Practice*, 46(8), 1252-1264.
- de Abreu e Silva, J. (2014). Spatial self-selection in land-use–travel behavior interactions: accounting simultaneously for attitudes and socioeconomic characteristics. *Journal of Transport and Land Use*, 7(2), 63-84.
- de Vos, J., & Witlox, F. (2016). Do people live in urban neighbourhoods because they do not like to travel? Analysing an alternative residential self-selection hypothesis. *Travel Behaviour and Society*, 4, 29-39.

- Drabas, T., & Wu, C. L. (2013). Modelling air carrier choices with a Segment Specific Cross Nested Logit model. *Journal of Air Transport Management*, 32, 8-16.
- Eluru, N., Bagheri, M., Miranda-Moreno, L. F., & Fu, L. (2012a). A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. *Accident Analysis & Prevention*, 47, 119-127.
- Eluru, N., Chakour, V., & El-Geneidy, A. M. (2012b). Travel mode choice and transit route choice behavior in Montreal: insights from McGill University members commute patterns. *Public Transport*, 4(2), 129-149.
- Ermagun, A., Hossein Rashidi, T., & Samimi, A. (2015). A joint model for mode choice and escort decisions of school trips. *Transportmetrica A: Transport Science*, 11(3), 270-289.
- Greene, W. H., & Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8), 681-698.
- Greene, W. H., & Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge University Press.
- Greene, W. H., & Hensher, D. A. (2013). Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. *Applied Economics*, 45(14), 1897-1902.
- Guerra, E. (2015). The geography of car ownership in Mexico City: a joint model of households' residential location and car ownership decisions. *Journal of Transport Geography*, 43, 171-180.
- Harding, C., Patterson, Z., Miranda-Moreno, L. F., & Zahabi, S. A. (2014). A spatial and temporal comparative analysis of the effects of land-use clusters on activity spaces in three Quebec cities. *Environment and Planning B: Planning and Design*, 41(6), 1044-1062.
- Hastings, D. (2003). Jobs densities for local areas: a new indicator. *Labour Market Trends*, 111(8), 407-414.
- He, X., & Zhang, L. (2014). Quantifying the Self-Selection Effects in Residential Location Choice with a Structural Equation Model. *Transportation Research Record: Journal of the Transportation Research Board*, 2453, 153-161.
- Innocenti, A., Lattarulo, P., & Pazienza, M. G. (2013). Car stickiness: Heuristics and biases in travel choice. *Transport Policy*, 25, 158-168.
- Jacques, C., & El-Geneidy, A. (2014). Does travel behavior matter in defining urban form? A quantitative analysis characterizing distinct areas within a region. *Journal of Transport and Land Use*, 7(1), 1-14.
- Jacques, C., Manaugh, K., & El-Geneidy, A. M. (2013). Rescuing the captive [mode] user: An alternative approach to transport market segmentation. *Transportation*, 40(3), 625-645.
- Kowald, M., Kieser, B., Mathys, N., & Justen, A. (2016). Determinants of mobility resource ownership in Switzerland: changes between 2000 and 2010. *Transportation*.

- Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological methods & research*, 33(2), 188-229.
- Larsen, J., El-Geneidy, A., & Yasmin, F. (2010). Beyond the Quarter Mile: Re-Examining Travel Distances by Active Transportation. *Canadian Journal Urban Research: Canadian Planning and Policy Supplement*, 19(1), 70-88.
- Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica: Journal of the Econometric Society*, 51(2), 507-512.
- Li, J., Walker, J., Srinivasan, S., & Anderson, W. (2010). Modeling private car ownership in China. *Transportation Research Record: Journal of the Transportation Research Board*, 2193, 76-84.
- Lin, J., & Long, L. (2008). What neighborhood are you in? Empirical findings of relationships between household travel and neighborhood characteristics. *Transportation*, 35(6), 739-758.
- Ma, L., & Srinivasan, S. (2010). Impact of individuals' immigrant status on household auto ownership. *Transportation Research Record: Journal of the Transportation Research Board*, 2156, 36-46.
- Magidson, J., & Vermunt, J. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, 20(1), 36-43.
- Manaugh, K., Miranda-Moreno, L. F., & El-Geneidy, A. M. (2010). The effect of neighbourhood characteristics, accessibility, home-work location, and demographics on commuting distances. *Transportation*, 37(4), 627-646.
- Miranda-Moreno, L. F., Eluru, N., Lee-Gosselin, M., & Kreider, T. (2012). Impact of ICT access on personal activity space and greenhouse gas production: evidence from Quebec City, Canada. *Transportation*, 39(5), 895-918.
- Nolan, A. (2010). A dynamic analysis of household car ownership. *Transportation Research Part A: Policy and Practice*, 44(6), 446-455.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling* 14 (4), 535-569.
- Paleti, R., Bhat, C., & Pendyala, R. (2013). Integrated model of residential location, work location, vehicle ownership, and commute tour characteristics. *Transportation Research Record*, 2382, 162-172.
- Patterson, Z., Saddier, S., Rezaei, A., & Manaugh, K. (2014). Use of the Urban Core Index to analyze residential mobility: the case of seniors in Canadian metropolitan regions. *Journal of Transport Geography*, 41, 116-125.
- Pinjari, A., Eluru, N., Bhat, C., Pendyala, R., & Spissu, E. (2008). Joint Model of Choice of Residential Neighborhood and Bicycle Ownership: Accounting for Self-Selection and Unobserved Heterogeneity. *Transportation Research Record: Journal of the Transportation Research Board*, 2082, 17-26.

- Pinjari, A., Pendyala, R., Bhat, C., & Waddell, P. (2007). Modeling residential sorting effects to understand the impact of the built environment on commute mode choice. *Transportation*, 34(5), 557-573.
- Pinjari, A. R., Pendyala, R. M., Bhat, C. R., & Waddell, P. A. (2011). Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, 38(6), 933-958.
- Portoghese, A., Spissu, E., Bhat, C. R., Eluru, N., & Meloni, I. (2011). A copula-based joint model of commute mode choice and number of non-work stops during the commute. *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, 337-362.
- Potoglou, D., & Kanaroglou, P. S. (2008). Modelling car ownership in urban areas: a case study of Hamilton, Canada. *Journal of Transport Geography*, 16(1), 42-54.
- Prillwitz, J., Harms, S., & Lanzendorf, M. (2006). Impact of life-course events on car ownership. *Transportation Research Record: Journal of the Transportation Research Board*, 1985, 71-77.
- Rana, T., Sikder, S., & Pinjari, A. (2010). Copula-based method for addressing endogeneity in models of severity of traffic crash injuries: application to two-vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, (2147), 75-87.
- Rashidi, T. H., & Mohammadian, K. (2016). Application of a nested trivariate copula structure in a competing duration hazard-based vehicle transaction decision model. *Transportmetrica A: Transport Science*, 1-18.
- Rashidi, T. H., Mohammadian, A., & Koppelman, F. S. (2011). Modeling interdependencies between vehicle transaction, residential relocation and job change. *Transportation*, 38(6), 909-932.
- Salon, D. (2009). Neighborhoods, cars, and commuting in New York City: A discrete choice approach. *Transportation Research Part A: Policy and Practice*, 43(2), 180-196.
- Salon, D. (2015). Heterogeneity in the relationship between the built environment and driving: Focus on neighborhood type and travel purpose. *Research in Transportation Economics*, 52, 34-45.
- Sanko, N., Dissanayake, D., Kurauchi, S., Maesoba, H., Yamamoto, T., & Morikawa, T. (2014). Household car and motorcycle ownership in Bangkok and Kuala Lumpur in comparison with Nagoya. *Transportmetrica A: Transport Science*, 10(3), 187-213.
- Schimek, P. (1996). Household Motor Vehicle Ownership and Use: How Much Does Residential Density Matter? *Transportation Research Record: Journal of the Transportation Research Board*, 1552, 120-125.
- Scott, D. M., & Axhausen, K. W. (2006). Household mobility tool ownership: modeling interactions between cars and season tickets. *Transportation*, 33(4), 311-328.



- Senbil, M., Kitamura, R., & Mohamad, J. (2009). Residential location, vehicle ownership and travel in Asia: A comparative analysis of Kei-Han-Shin and Kuala Lumpur metropolitan areas. *Transportation*, 36(3), 325-350.
- Shen, J. (2009). Latent class model or mixed logit model? A comparison by transport mode choice data. *Applied Economics*, 41(22), 2915-2924.
- Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6), 449-460.
- Sobhani, A., Eluru, N., & Faghih-Imani, A. (2013). A latent segmentation based multiple discrete continuous extreme value model. *Transportation Research Part B: Methodological*, 58, 154-169.
- Song, Y., & Knaap, G.-J. (2007). Quantitative classification of neighbourhoods: The neighbourhoods of new single-family homes in the Portland Metropolitan Area. *Journal of Urban Design*, 12(1), 1-24.
- Tang, W. L., & Mokhtarian, P. L. (2009). Accounting for taste heterogeneity in purchase channel intention modeling: An example from northern California for book purchases. *Journal of Choice Modelling*, 2(2), 148-172.
- Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.
- Van Acker, V., Mokhtarian, P. L., & Witlox, F. (2014). Car availability explained by the structural relationships between lifestyles, residential location, and underlying residential and travel attitudes. *Transport Policy*, 35, 88-99.
- Van Acker, V., & Witlox, F. (2010). Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship. *Journal of Transport Geography*, 18(1), 65-74.
- Van Wee, B. (2002). Land use and transport: research and policy challenges. *Journal of Transport Geography*, 10(4), 259-271.
- Vance, C., & Hedel, R. (2007). The impact of urban form on automobile travel: disentangling causation from correlation. *Transportation*, 34(5), 575-588.
- Waddell, P., Bhat, C., Eluru, N., Wang, L., & Pendyala, R. M. (2007) Modeling interdependence in household residence and workplace choices. *Transportation Research Record* (pp. 84-92).
- Wafa, Z., Bhat, C. R., Pendyala, R. M., & Garikapati, V. M. (2015). Latent-segmentation-based approach to investigating spatial transferability of activity-travel models. *Transportation Research Record: Journal of the Transportation Research Board*, 2493, 136-144.
- Walker, J. L., & Li, J. (2007). Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems*, 9(1), 77-101.
- Weinberger, R., & Goetzke, F. (2010). Unpacking preference: How previous experience affects auto ownership in the United States. *Urban Studies*, 47(10), 2111-2128.

- Wen, C. H., & Lai, S. C. (2010). Latent class models of international air carrier choice. *Transportation Research Part E: Logistics and Transportation Review*, 46(2), 211-221.
- Xie, Y., Zhao, K., & Huynh, N. (2012). Analysis of driver injury severity in rural single-vehicle crashes. *Accident Analysis & Prevention*, 47, 36-44.
- Yasmin, S., Eluru, N., Pinjari, A. R. & Tay, R. (2014). Examining driver injury severity in two vehicle crashes - A copula based approach. *Accident Analysis and Prevention*, 66, 120-135.
- Zou, Y., Zhang, Y., & Zhu, X. (2014). Constructing a bivariate distribution for freeway speed and headway data. *Transportmetrica A: Transport Science*, 10(3), 255-272.

**Figure 1: Residential Location Clusters**

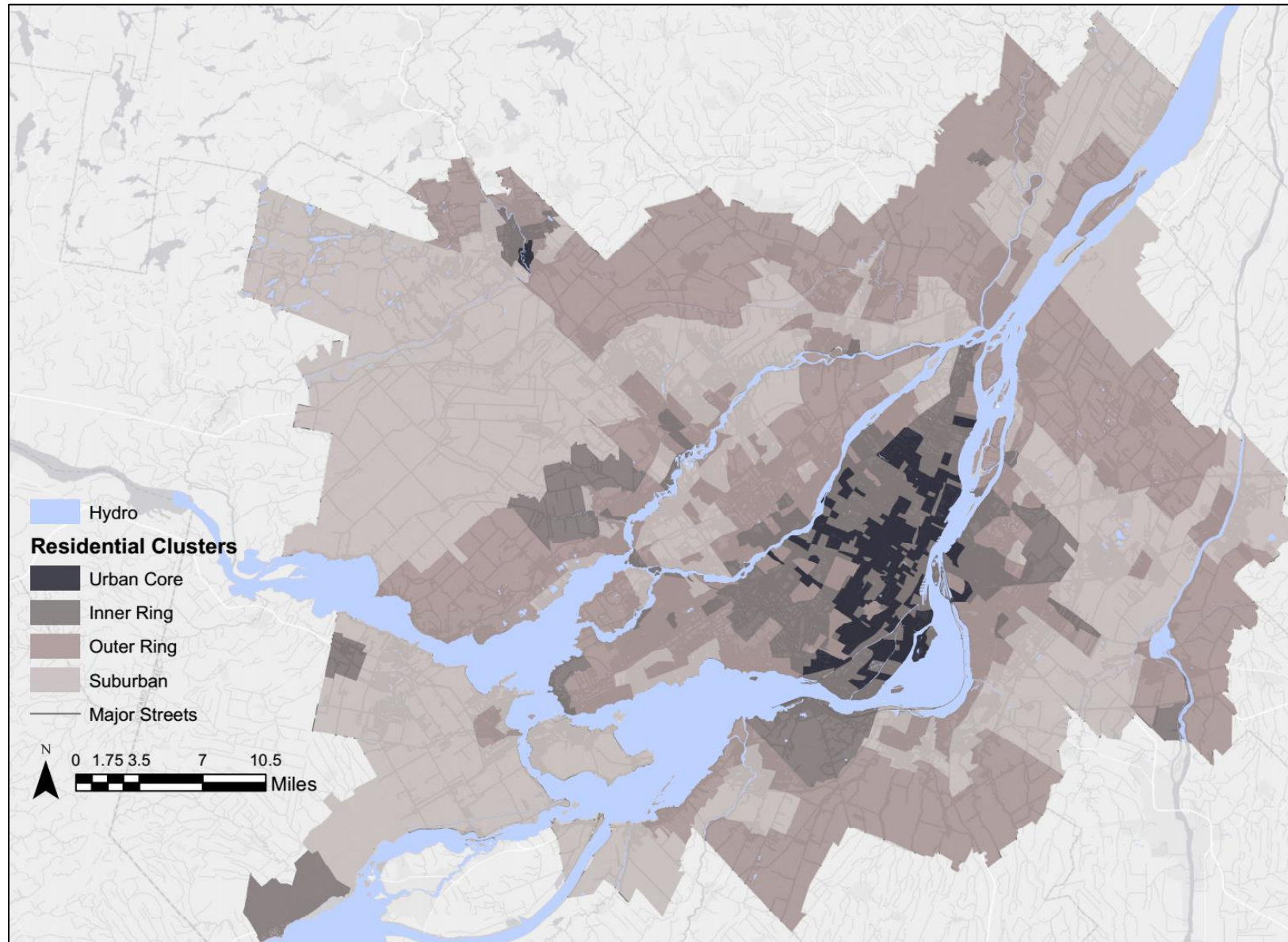
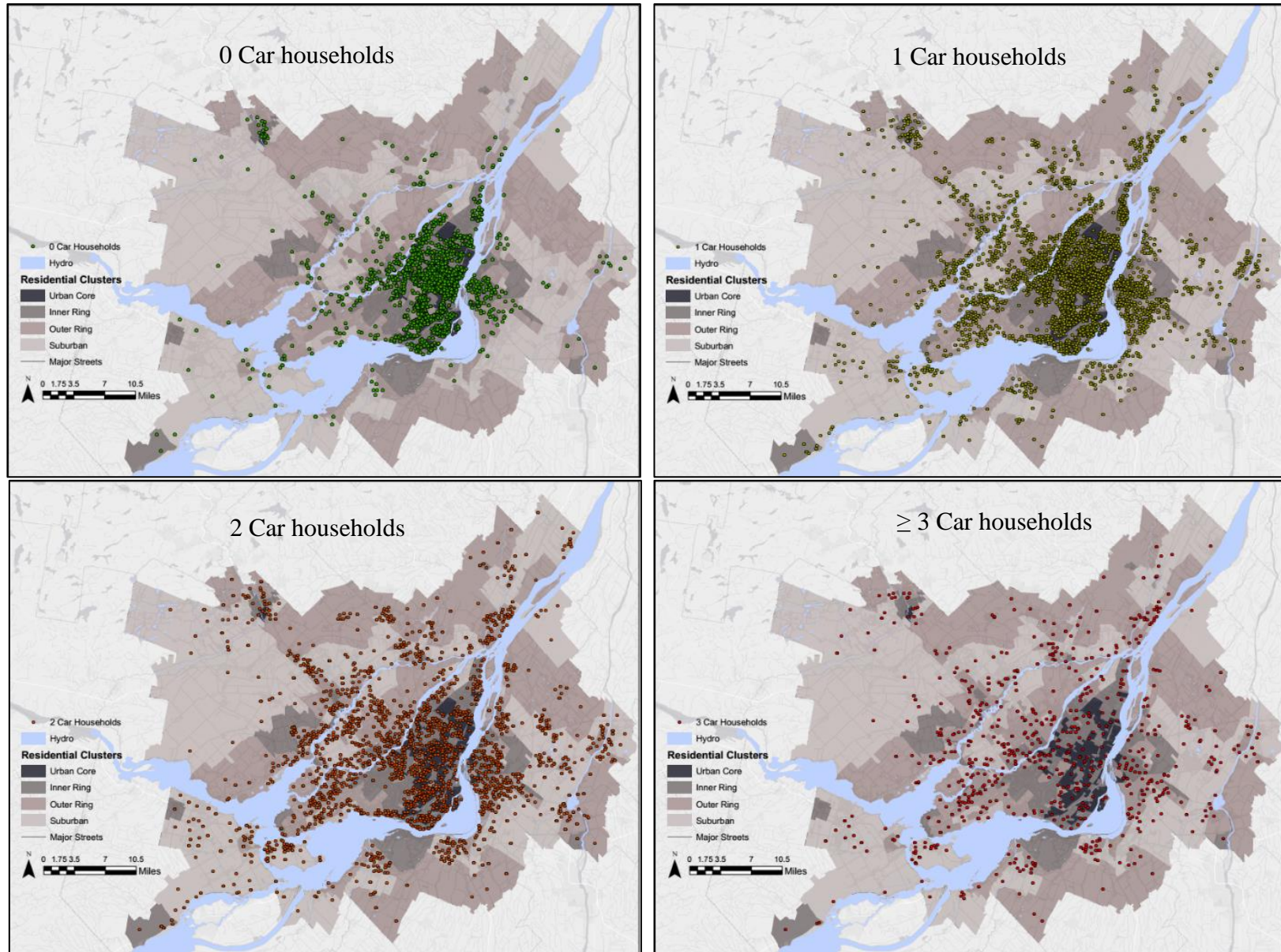




Figure 2: Vehicle Ownership Distribution across Residential Location Clusters



**Table 1 (a): Literature on residential self-selection bias in the context of vehicle ownership (explicitly considering residential location as a decision variable)**

	Study	Study Area	Methodology	Choice Dimensions for which Endogeneity is Investigated	Independent Variables			
					Household Demographics	Built Environment	Transit Accessibility	Policy Variables
1.	Guerra (2015)	Mexico City, Mexico	Mixed multinomial logit	Vehicle ownership and residential location	√	---	---	---
2.	Bhat et al. (2014)	San Francisco, Oakland, San Jose, USA	Multidimensional choice model	Vehicle ownership, residential location, number of motorized and non-motorized tours, and average tour distance	√	√	---	---
3.	He and Zhang (2014)	Washington DC, USA	Structural equation model	Vehicle ownership, residential location, and vehicle mileage	√	√	---	---
4.	de Abreu e Silva (2014)	Lisbon, Portugal	Structural equation model	Vehicle ownership, commute distance, number of trips, and time between first and last trips	√	√	√	---
5.	Paleti et al. (2013)	San Francisco Bay, USA	Mixed Multidimensional choice model	Vehicle ownership, residential location, work location, and commute tour characteristics	√	√	---	---
6.	Bhat et al. (2013)	San Francisco Bay, USA	Bivariate multinomial probit	Vehicle ownership and residential location	√	---	---	---
7.	Pinjari et al. (2011)	San Francisco Bay, USA	Mixed multidimensional choice model	Vehicle ownership, residential location, bicycle ownership, and commute tour mode	√	√	√	√
8.	Weinberger and Goetzke (2010)	Boston, Chicago, Philadelphia, San Francisco, Washington DC, USA	Multinomial probit	Vehicle ownership and residential location	√	√	---	---

9.	Salon (2009)	New York City, USA	Multinomial logit	Vehicle ownership, residential location and commute transport mode	√	√	√	√
10.	Senbil et al. (2009)	Kei-Han-Shin, Japan; Kuala Lumpur, Malaysia	Structural equation model	Vehicle ownership, residential location, and vehicle use	√	√	√	---
11.	Bhat and Guo (2007)	San Francisco Bay, USA	Mixed multidimensional choice model	Vehicle ownership and residential location	√	√	√	√

**Table 1 (b): Literature on residential self-selection bias in the context of vehicle ownership (only considering the interaction of built environment and vehicle ownership)**

	Study	Study Area	Methodology	Choice Dimensions for which Endogeneity is Investigated	Independent Variables			
					Household Demographic	Built Environment	Accessibility	Attitude/Lifestyle
1.	van Acker et al. (2014)	Flanders, Belgium	Structural equation model	Vehicle availability and residential land use characteristics	√	√	---	√
2.	Cao and Cao (2014)	Minneapolis-St. Paul, USA	Ordered logit and Statistical control approach	Vehicle ownership, attitude and preference, and transit oriented development	√	√	√	---
3.	Brownstone and Fang (2014)	USA	Bayesian multivariate ordered probit	Vehicle ownership, usage and residential density	√	---	√	---
4.	de Abreu e Silva (2014)	Lisbon, Portugal	Structural equation model	Vehicle ownership, residence and workplace land use, commuting distance, trip scheduling, and number of trips	√	√	√	---
5.	Aditjandra et al. (2012)	Tyne and Wear, UK	Dynamic structural equation model	Changes in vehicle ownership and changes in driving behavior	√	√	---	√
6.	de Abreu e Silva et al. (2012a)	Greater Montreal Area, Canada	Structural equation model	Vehicle ownership, residence and workplace land use, commuting distance, trip scheduling, and number of trips	√	√	√	---
7.	de Abreu e Silva et al. (2012b)	California, USA	Structural equation model	Vehicle ownership, residence and workplace land use, commuting distance, trip scheduling, and number of trips	√	√	√	---
8.	van Acker and Witlox (2010)	Ghent, Belgium	Structural equation model	Vehicle ownership and vehicle use	√	√	√	---

9.	de Abreu e Silva and Goluias (2009)	Lisbon, Portugal	Structural equation model	Residential location, workplace location, commuting distance and vehicle ownership	√	√	√	---
10.	Chen et al. (2008)	New York, USA	Simultaneous equation model	Vehicle ownership and vehicle use	√	√	√	---
11.	Beckman et al. (2008)	California, USA	Latent class cluster analysis	Vehicle ownership, residential location, immigration and commuting behavior	√	√	---	---
12.	Cao et al. (2007)	Northern California, USA	Structural equation model	Changes in spaciousness, accessibility, attractiveness, driving behavior, vehicle ownership, and walking behavior	√	√	---	√
13.	de Abreu e Silva et al. (2006)	Lisbon, Portugal	Structural equation model	Residential location, workplace location, commuting distance and vehicle ownership	√	√	√	---



**Table 2: Average Characteristics of Census Tracts by Residential Neighborhood Types**

<b>Attributes</b>	<b>All</b>	<b>Urban Core</b>	<b>Inner Suburb</b>	<b>Outer Suburb</b>	<b>Suburban</b>
Population density (per ha)	22.79	43.74	24.23	9.91	5.60
Job density	1.13	0.96	2.08	0.54	0.35
# of detached households	638.73	83.16	183.42	1061.01	1988.25
# of transit, walk, and bike commuters	558.64	942.69	448.42	393.88	437.32
# of dwellings built before 1946	215.95	544.93	124.74	80.99	89.86
<b>Number of observations</b>	878	226	289	251	112

**Table 3(a): Sample Characteristics – Land Use and Transit Attributes**

Variable	Definition	CT Level Minimum	CT Level Maximum	CT level Mean
Connection ratio	Ratio of # of intersections in CT to sum of # of intersections and # of dead ends in CT	0	1.000	0.950
Area	Ln (total area of CT, Acre)	2.245	11.127	5.635
Dwelling density	Ln (ratio of # of dwellings to residential area in CT)	-0.257	18.839	2.514
Population density	Ratio of total population to total area of CT in ha	0	183.467	23.250
Job density	Ratio of # of jobs to # of working age (15-65 years) population in CT	0	143.363	1.067
# of schools	# of schools in CT	0	10.000	1.876
# of markets	# of markets in CT	0	43.000	2.144
# of rented households	Ln (# of rented households in CT)	0	8.213	6.279
# of detached households	Ln (# of detached households in CT)	0	8.372	4.958
# of low-rise apartments	Ln (# of low-rise apartments in CT)	0	8.245	6.001
Driver commuters	Ln (# of driver commuters in CT)	0	8.588	6.853
Transit commuters	Ln (# of transit commuters in CT)	0	7.467	5.748
Land use mix	$-\sum_k \frac{[p_k \ln p_k]}{\ln(K)}$ , $p_k$ = proportion of developed land in the $k^{\text{th}}$ land use (residential, commercial, industrial, institutional and park facilities). It varies between 0 - 1; 0 = homogenous, 1 = perfectly heterogeneous mix	0	0.870	0.447
Bus destination diversity	No of bus routes operating in a CT	0	122.000	5.872

Variable		Frequency (%)		
Median Income level of CT				
Low income (< 40 K)		196 (22.8)		
Medium income (40K-80K)		529 (61.6)		
High income (> 80K)		134 (15.6)		
Variable Names	Definition	Household Level Minimum	Household Level Maximum	Household level Mean
Bus stops	# of bus stops in 600m buffer	0	92.000	24.401
Commuter rail stops	# of commuter rail stations in 600m buffer	0	2.000	0.041
Bike route length	Ln (bike route length in meters in 600m buffer)	0	8.734	2.472
Pedestrian/bike street	Ln (length of pedestrian/bike street in 600m buffer)	0	10.173	9.354
Building footprint	Ln (building footprint in square meters in 600m buffer)	0	13.009	10.150
Distance to CBD	Ln (distance to central business district in meter from household)	4.753	11.069	9.367

**Table 3(b): Sample Characteristics – Household Characteristics**

	Residential Location Alternatives				Vehicle Ownership Levels			
	UC	IR	OR	SU	0	1	2	3+
Overall share	<b>26.8%</b>	<b>25.1%</b>	<b>28.8%</b>	<b>19.3%</b>	<b>20.6%</b>	<b>43.1%</b>	<b>29.1%</b>	<b>7.2%</b>
<u>Household size</u>								
1 person	37.4%	30.7%	20.9%	11.0%	46.2%	49.5%	3.0%	1.3%
2 persons	25.1%	25.9%	29.9%	19.1%	14.2%	48.6%	34.0%	3.2%
≥ 3 persons	19.7%	19.9%	34.3%	26.1%	6.2%	32.4%	45.3%	16.1%
<u>Presence of children</u>								
Kids 0-4	23.0%	22.6%	28.0%	26.5%	8.9%	37.2%	49.7%	4.1%
Kids 5-9	21.3%	19.6%	32.3%	26.8%	7.1%	37.8%	49.8%	5.4%
Kids 10-14	19.3%	17.9%	35.9%	26.9%	6.8%	35.9%	47.7%	9.6%
Kids 15-18	18.9%	20.1%	35.4%	25.6%	7.8%	33.2%	40.7%	18.4%
<u># of workers</u>								
Full-time								
0	29.6%	30.9%	26.2%	13.2%	37.4%	49.4%	11.3%	1.8%
1	29.1%	24.3%	27.0%	19.5%	17.3%	48.7%	28.6%	5.5%
≥ 2	20.6%	19.1%	34.1%	26.2%	4.1%	29.3%	50.8%	15.8%
Part-time								
0	26.6%	25.5%	28.7%	19.2%	21.2%	43.8%	28.3%	6.8%
1	28.4%	21.7%	30.2%	19.7%	15.4%	37.3%	36.4%	10.9%
≥ 2	26.7%	25.0%	30.0%	18.3%	11.7%	41.6%	31.7%	15.0%
<u># of students</u>								
0	28.3%	27.0%	27.1%	17.5%	24.9%	46.5%	23.9%	4.7%
1	27.5%	23.4%	29.0%	20.1%	14.4%	37.7%	36.2%	11.7%
≥ 2	19.2%	18.3%	36.4%	26.1%	7.4%	33.6%	45.1%	13.9%
<u># of retirees</u>								
0	27.8%	23.1%	28.5%	20.6%	15.9%	40.5%	34.8%	8.9%
1	27.6%	30.1%	26.9%	15.3%	38.9%	41.4%	14.9%	4.8%
≥ 2	19.6%	27.6%	34.0%	8.8%	13.7%	60.7%	22.9%	2.7%
<u>Median Income</u>								
Low (<40K)	63.3%	36.7%	0.0%	0.0%	44.5%	43.0%	10.7%	1.8%
Medium (40K-80K)	21.1%	27.5%	35.4%	16.0%	17.7%	46.0%	29.1%	7.2%
High (>80K)	26.8%	25.1%	28.8%	19.3%	3.9%	32.7%	49.9%	13.5%

UC= Urban Core; IR = Inner Ring; OR = Outer Ring; SU = Suburban

**Table 4: MOL Estimation Results**

<b>Variables</b>	<b>Estimate</b>	<b>t-stat</b>
Thresholds		
Threshold 1	0.422	1.360
Threshold 2	3.674	11.726
Threshold 3	6.718	21.197
<b>Household Socio-Demographic Characteristics</b>		
# of household members	0.109	3.224
# of male adults	1.098	20.240
# of female adults	0.614	10.987
Presence of children (0-4 years)	0.205	2.091
Presence of children (15-18 years)	0.223	2.843
# of adults (35-64 years)	0.319	10.065
Full-time workers	0.818	19.033
Part-time workers	0.533	8.054
# of retirees	0.178	4.110
# of workers with no bus transit access at workplace	0.322	2.805
# of workers with low bus transit access at workplace	0.213	5.062
# of workers with high bust transit access at workplace	-0.234	-4.625
<b>Land Use and Built Environment Characteristics</b>		
<i>Median Income Level of CT (Base: Low income)</i>		
Medium income (40-80K)	0.290	4.539
High income (>80K)	0.463	4.461

<i>Standard deviation</i>	0.077	4.892
Connection ratio	-0.922	-2.697
# of schools	-0.039	-2.561
# of markets	-0.030	-3.358
# of rented households	-0.157	-4.475
Bike route length	-0.018	-2.600
Driver commuters	0.328	6.502
Transit commuters	-0.183	-5.230
Dwelling density	-0.069	-2.670
# of detached households	0.089	5.141
<b>Transit Accessibility Measures</b>		
Bus stops	-0.010	-4.089
Log-likelihood at constants (N=10,214)	-12637.56	
Log-likelihood at convergence (N=10,214)	-9096.12	

**Table 5: ESOL Model Estimates**

<b>Variables</b>	<b>Urban Core</b>	<b>Inner Ring</b>	<b>Outer Ring</b>	<b>Suburban</b>
<b>Thresholds</b>				
Threshold 1	0.277 (0.329) †	-0.321 (-0.965)	-1.995 (-6.215)	1.596 (1.079)
Threshold 2	3.309 (3.922)	2.730 (8.057)	1.425 (4.479)	5.301 (3.572)
Threshold 3	5.707 (6.662)	5.304 (14.925)	4.571 (13.872)	8.789 (5.884)
<b>Household Socio-Demographic Characteristics</b>				
# of household members	–	0.263 (5.924)	0.743 (7.122)	0.683 (5.267)
Single person (Base: Multi-person)	-0.511 (-3.883)	–	–	–
Presence of children (0-4 years)	0.474 (3.165)	–	-0.895 (-4.503)	-0.757 (-3.143)
# of male adults	0.852 (8.060)	0.694 (8.356)	–	–
# of female adults	0.517 (4.761)	–	–	–
Full-time workers	0.767 (12.504)	0.986 (12.867)	0.645 (6.783)	0.911 (8.017)
Part-time workers	0.240 (1.996)	0.589 (4.316)	0.731 (5.267)	0.659 (3.813)
# of young adults (19-34 years)	-0.453 (-7.280)	–	0.683 (8.706)	0.828 (7.846)
# of middle aged adults (35-64 years)	–	0.303 (4.968)	0.753 (10.230)	0.901 (9.202)
# of students	–	–	-0.456 (-4.222)	-0.432 (-3.225)
# of retirees	–	0.461 (6.249)	0.362 (3.510)	0.327 (2.576)
# of workers with bus, metro and commuter train access at workplace	–	-0.299 (-2.045)	–	–
# of workers with no bus transit access at workplace	–	0.579 (2.038)	–	–
# of workers with low bus transit access at workplace	0.566 (5.812)	–	0.0661 (3.381)	–
# of workers with high bust transit access at workplace	–	–	–	-0.313 (-2.665)

<b>Land Use and Built Environment Characteristics</b>				
Population density	–	–	-0.023 (-2.461)	–
Job density	–	–	-0.086 (-2.081)	–
Driver commuters	0.483 (5.181)	0.734 (8.053)	–	0.414 (2.171)
Transit commuters	–	-0.256 (-3.326)	-0.224 (-4.077)	–
# of single detached households	–	0.083 (2.985)	–	–
# of low-rise apartments	-0.285 (-4.195)	-0.124 (-2.894)	–	–
Dwelling density	-0.466 (-4.727)	–	-0.302 (-3.131)	–
# of rented dwellings	–	-0.305 (-3.304)	–	-0.201 (-3.416)
Pedestrian/bike street	–	-0.229 (-4.881)	–	–
<i>Median Income Level of CT (Base: Low income)</i>				
Medium income (40-80K)	–	–	-0.227 (-2.410)	–
# of schools	-0.051 (-2.226)	–	–	–
# of bars	–	–	-0.164 (-2.888)	–
<b>Transit Accessibility Measures</b>				
Bus destination diversity	–	-0.014 (-2.366)	–	–
Bus stops	–	–	–	-0.036 (-6.272)
Commuter rail stops	–	–	-0.409 (-2.461)	–
# of observations	2734	2567	2946	1967
Log-likelihood at constants	-2895.23	-2954.16	-3495.51	-2191.04
Log-likelihood at convergence	-2403.54	-2363.21	-2628.77	-1636.67

† The values in the parenthesis are *t*-stats



**Table 6: LSOLII Estimation Results**

<b>Segmentation Component</b>				
<b>Variables</b>	<b>Segment 1</b>		<b>Segment 2</b>	
	<b>Estimate</b>	<b>t-stat</b>	<b>Estimate</b>	<b>t-stat</b>
Constants	–	–	1.516	3.183
Area	–	–	-0.197	-4.045
Land use mix	–	–	0.748	2.584
# of detached households	–	–	-0.191	-4.756
# of rented households	–	–	0.132	2.494
<i>Residential Cluster (Base: Inner and outer ring)</i>				
Urban core	–	–	0.457	2.626
<b>Vehicle Ownership Component</b>				
Thresholds				
Threshold 1	-3.379	-2.097	0.712	1.564
Threshold 2	1.133	0.730	4.108	8.882
Threshold 3	7.066	4.708	6.244	13.595
<b>Household Socio-Demographic Characteristics</b>				
# of household members	–	–	0.227	7.565
Presence of children (15-18 years)	1.181	4.945	–	–
# of male adults	3.863	12.475	0.415	6.339
# of female adults	2.837	10.839	–	–
Full-time workers	0.721	4.888	0.871	13.981
Part-time workers	0.824	3.511	0.443	4.475
# of adults (19-24 years)	–	–	-0.196	-2.553
# of adults (25-34 years)	0.662	3.762	0.309	6.029

# of adults (35-64 years)	0.644	5.346	–	–
# of retirees	-0.396	-2.680	0.353	5.683
# of workers with bus, metro and commuter train access at workplace	-0.831	-3.344	–	–
# of workers with no bus transit access at workplace	–	–	0.389	2.111
# of workers with high bus transit access at workplace	–	–	-0.191	-2.488
# of workers with low bus transit access at workplace	–	–	0.248	4.012
<b>Land Use and Built Environment Characteristics</b>				
Driver commuters	–	–	0.446	7.166
Transit commuters	–	–	-0.307	-5.493
Dwelling density	–	–	-0.307	-5.259
# of low-rise apartments	-0.128	-2.871	–	–
Bike route length	-0.051	-2.135	–	–
Building footprint	–	–	-0.052	-2.887
Distance to CBD	-0.441	-2.823	–	–
# of schools	–	–	-0.061	-3.053
<i>Median Income Level of CT (Base: Low income)</i>				
Medium income (40-80K)	–	–	0.344	4.262
High income (>80K)	–	–	0.670	4.965
<b>Transit Accessibility Measures</b>				
Bus stops	-0.047	-5.442	–	–
Log-likelihood at constants (N=10,214)	-12637.56			
Log-likelihood at convergence (N=10,214)	-8858.49			

**Table 7(a): Copula MNL (Residential Cluster Choice) Model Estimation Results**

Variables	Urban Core	Inner Ring	Outer Ring	Suburban
Constant	–	-0.213 (-4.775) <sup>†</sup>	-1.190 (-19.460)	-1.778 (-24.050)
# of children	–	–	0.094 (3.909)	0.391 (8.753)
# of male adults	–	–	-0.269 (-6.293)	-0.269 (-6.293)
# of young adults (19-34 years)	–	-0.143 (-4.417)	-0.384 (-10.045)	-0.472 (-11.002)
# of students	–	–	–	-0.349 (-7.226)
# of retirees	–	0.218 (7.286)	0.218 (7.286)	–
# of license holders	–	0.074 (2.732)	1.141 (31.147)	1.359 (30.588)
# of workers with no bus transit access at workplace	–	–	0.634 (6.574)	0.634 (6.574)
# of workers with low bus transit access at workplace	–	0.396 (9.605)	0.396 (9.605)	0.396 (9.605)
# of workers with medium bus transit access at workplace	–	–	-0.379 (-8.864)	-0.548 (-10.481)
# of workers with high bus transit access at workplace	–	-0.152 (-3.144)	-0.644 (-10.794)	-0.812 (-12.091)
# of part-time workers	0.228 (3.274)	–	–	–

<sup>†</sup>The values in the parenthesis are *t*-stats

**Table 7(b): Copula OL (Vehicle Ownership) Estimation Results**

<b>Variables</b>	<b>Urban Core</b>	<b>Inner Ring</b>	<b>Outer Ring</b>	<b>Suburban</b>
<b>Thresholds</b>				
Threshold 1	1.849 (2.882) †	1.348 (5.772)	-2.900 (-9.741)	0.218 (0.159)
Threshold 2	4.293 (6.690)	3.570 (14.934)	0.245 (0.839)	3.582 (2.622)
Threshold 3	6.438 (9.904)	5.716 (21.730)	3.118 (10.514)	6.705 (4.869)
<b>Household Socio-Demographic Characteristics</b>				
Single person household (Base: Multi-person)	-0.237 (-2.861)	–	–	–
Presence of children (0-4 years)	0.238 (2.130)	–	–	–
# of male adults	0.326 (5.357)	0.491 (8.325)	–	–
# of full-time workers	0.655 (13.296)	0.656 (12.134)	0.995 (15.189)	1.009 (12.188)
# of part-time workers	0.441 (4.187)	0.218 (2.265)	1.075 (9.071)	0.843 (6.225)
# of adults (19-24 years)	–	–	0.842 (13.215)	1.008 (13.044)
# of adults (25-64 years)	–	0.111 (2.464)	0.865 (14.668)	1.001 (13.520)
# of students	–	–	0.196 (4.613)	0.177 (3.300)
# of retirees	–	0.294 (5.313)	0.731 (11.051)	0.665 (8.351)
<b>Land Use and Built Environment Characteristics</b>				
Population density	–	–	-0.020 (-2.309)	–
Job density	–	–	-0.074 (-2.562)	–
# of driver commuters	0.336 (4.635)	0.459 (7.054)	–	0.360 (2.060)
# of transit commuters	–	-0.146 (-2.684)	-0.210 (-4.143)	–
# of detached households	–	0.059 (3.097)	–	–

# of low-rise apartments	-0.165 (-3.355)	-0.090 (-2.930)	–	–
Dwelling density	-0.299 (-3.996)	–	-0.288 (-2.893)	–
# of rented dwellings	–	-0.161 (-2.521)	–	-0.175 (-3.262)
Pedestrian/bike street	–	-0.153 (-4.746)	–	–
<i>Median Income Level of CT (Base: Low income)</i>				
Medium income (40-80K)	–	–	-0.180 (-1.968)	–
# of schools	-0.038 (-2.208)	–	–	–
# of bars	–	–	-0.147 (-2.775)	–
<b>Transit Accessibility Measures</b>				
Bus destination diversity	–	-0.009 (-2.204)	–	–
# of bus stops	–	–	–	-0.034 (-6.143)
# of commuter rail stops	–	–	-0.366 (-2.344)	–
<b>Copula Parameters</b>				
Copula	<b>Joe</b>	<b>Joe</b>	<b>FGM</b>	<b>FGM</b>
Correlation parameters	3.521 (13.199)	3.903 (13.052)	-1.000	-1.000

† The values in the parenthesis are *t*-stats

**Table 8: Prediction Comparison (Estimation Sample, N = 10,214)**

Disaggregate Level						
Summary statistic		MOL	ESOL	LSOLII	Copula OL (Joe-FGM)	
# of parameters		28	67	42	55	
Log-likelihood at constants		-12637.56	-12637.56	-12637.56	-12637.56	
Log-likelihood at convergence		-9096.12	-9075.30	-8860.60	<b>-8481.99</b>	
BIC		18451	18769	18081	<b>17472</b>	
Aggregate Level						
Vehicle Ownership Levels/Measures of Fit		Actual shares	MOL	ESOL	LSOLII	Copula OL (Joe-FGM)
0 Car		20.6	20.5	20.0	20.6	19.8
1 Car		43.1	43.5	42.9	43.3	43.9
2 Cars		29.1	28.7	28.7	28.7	28.8
≥ 3 Cars		7.2	7.3	8.4	7.4	7.5
RMSE		–	0.29	0.71	<b>0.24</b>	0.60
MAPE		–	<b>0.01</b>	0.05	<b>0.01</b>	0.03
No full time worker	Log-likelihood	–	-3166.40	-3120.88	-3051.15	<b>-2856.97</b>
Full-time worker present	Log-likelihood	–	-6040.35	-5954.41	-5807.33	<b>-5625.02</b>
Urban core	Log-likelihood	–	-2542.93	-2414.98	-2400.92	<b>-2258.88</b>
Medium income	Log-likelihood	–	-5832.26	-5370.46	-5626.92	<b>-5370.46</b>
No bus stops	Log-likelihood	–	-575.01	-591.54	<b>-543.67</b>	-565.58
Bus stops present	Log-likelihood	–	-8631.75	-8483.76	-8314.81	<b>-7916.41</b>

**Table 9: Prediction Comparison (Validation Sample)**

Disaggregate Level					
Summary statistic		MOL	ESOL	LSOLII	Copula OL (Joe-FGM)
# of parameters		28	67	42	55
Log-likelihood at constants		-2154.2 (-2167.8/-2140.6) ††	-2154.2 (-2167.8/-2018.3)	-2154.2 (-2167.8/-2018.3)	-2154.2 (-2167.8/-2018.3)
Predictive log-likelihood		1565.1 (-1576.5/-1553.6)	1543.3 (-1554.6/-1532.1)	-1531.05 (-1541.9/-1520.2)	<b>-1441.7</b> <b>(-1452.3/-1431.1)</b>
BIC		3339 (3316/3362)	3586 (3564/3609)	3375 (3353/3397)	<b>3294</b> <b>(3272/3315)</b>
Aggregate Level					
Vehicle Ownership Levels/Measures of Fit		MOL	ESOL	LSOLII	Copula OL (Joe-FGM)
0 Car		20.1 (20.0/20.3)	19.6 (19.5/19.7)	20.5 (20.3/20.6)	19.4 (19.2/19.5)
1 Car		43.5 (43.3/43.6)	42.9 (42.8/43.0)	43.3 (43.2/43.5)	44.0 (43.9/44.1)
2 Cars		29.2 (29.1/29.3)	29.3 (29.2/29.4)	28.8 (28.7/30.0)	29.3 (29.2/29.5)
≥ 3 Cars		7.2 (7.1/7.3)	8.2 (8.1/8.3)	7.4 (7.2/7.4)	7.3 (7.2/7.4)
RMSE		0.97 (0.88/1.06)	<b>0.76</b> <b>(0.67/0.85)</b>	1.11 (1.01/1.21)	0.86 (0.77/0.96)
MAPE		0.04 (0.04/0.05)	<b>0.04</b> <b>(0.03/0.04)</b>	0.05 (0.05/0.06)	<b>0.04</b> <b>(0.03/0.04)</b>
No full-time worker present	Predictive LL	-505.4 (-508.1/-496.7)	-500.5 (-506.3/-494.8)	-497.7 (-503.4/-492.0)	<b>-458.2</b> <b>(-463.6/-452.8)</b>
Full-time worker present	Predictive LL	-1062.6 (-1072.5/-1052.8)	-1042.8 (-1052.1/-1033.5)	-1033.4 (-1042.6/-1024.1)	<b>-983.5</b> <b>(-992.2/-974.7)</b>

Urban core	Predictive LL	-420.7 (-427.2/-414.2)	-402.8 (-408.8/-396.8)	-399.6 (-405.6/393.6)	<b>-377.0</b> <b>(-382.9/-371.1)</b>
Medium income	Predictive LL	-1022.7 (-1032.2/-1013.2)	-1007.9 (-1017.2/-998.6)	-1003.4 (-1012.3/-994.6)	<b>-943.0</b> <b>(-951.9/-934.2)</b>
No bus stops	Predictive LL	-101.7 (-103.7/-99.6)	-102.2 (-104.2/-100.1)	<b>-98.5</b> <b>(-100.5/-96.4)</b>	-99.0 (-101.0/-97.0)
Bus stops present	Predictive LL	-1463.4 (-1474.2/-1452.6)	-1441.2 (-1451.7/-1430.7)	-1432.6 (-1442.8/-1422.4)	<b>-1342.7</b> <b>(-1352.6/-1332.8)</b>

†† The values in the parenthesis are 95% confidence bands



**Table 10: Policy Analysis Results<sup>†</sup>**

Explanatory Variables	MOL		ESOL		LSOL		Copula OL	
	0 Car	≥ 3 Cars	0 Car	≥ 3 Cars	0 Car	≥ 3 Cars	0 Car	≥ 3 Cars
# of full-time workers increased by 1	-37.94 (-38.64/-37.24)	68.89 (67.76/70.01)	-38.62 (-39.90/-37.37)	62.13 (59.83/64.37)	-39.32 (-40.64/-38.01)	56.32 (53.99/58.65)	-44.55 (-45.68/-43.42)	85.22 (83.27/87.18)
# of part-time workers increased by 1	-26.26 (-27.46/-25.06)	42.54 (40.14/44.94)	-23.85 (-26.29/-21.47)	45.45 (40.41/50.36)	-25.08 (-27.34/-22.82)	37.06 (32.53/41.58)	-28.75 (-31.24/-26.26)	65.53 (60.30/70.76)
# of bus stops increased by 10%	1.49 (1.35/1.63)	-0.86 (-0.94/-0.77)	0.68 (0.62/0.74)	-0.62 (-0.65/0.59)	1.05 (0.97/1.14)	-1.07 (-1.13/-1.01)	0.76 (0.69/0.83)	-0.69 (-0.72/-0.66)
Bike route length increased by 10%	0.31 (0.26/0.36)	-0.13 (-0.15/-0.11)	0	0	0.11 (0.0/0.13)	-0.09 (-0.11/-0.07)	0	0
Pedestrian/bike street increased by 10%	0	0	3.49 (3.13/3.85)	-2.93 (-3.02/-2.83)	0	0	3.37 (3.04/3.70)	-2.16 (-2.22/-2.11)

<sup>†</sup>The values in the parenthesis are 95% confidence bands