A LATENT SEGMENTATION BASED MULTIPLE DISCRETE CONTINUOUS EXTREME VALUE MODEL

Anae Sobhani
PhD Student
Department of Civil Engineering and Applied Mechanics
McGill University
Ph: 647 894 2613, Fax: 514 398 7361
Email:anae.sobhani@mail.mcgill.ca

Naveen Eluru*
Assistant Professor
Department of Civil Engineering and Applied Mechanics
McGill University
Ph: 514 398 6823, Fax: 514 398 7361
Email:naveen.eluru@mcgill.ca

Ahmadreza Faghih-Imani
PhD Student
Department of Civil Engineering and Applied Mechanics
McGill University
Ph: 514-652-4464, Fax: 514-398-7361
Email: seyed.faghihimani@mail.mcgill.ca

*Corresponding Author

ABSTRACT

We examine an alternative method to incorporate potential presence of population heterogeneity within the Multiple Discrete Continuous Extreme Value (MDCEV) model structure. Towards this end, an endogenous segmentation approach is proposed that allocates decision makers probabilistically to various segments as a function of exogenous variables. Within each endogenously determined segment, a segment specific MDCEV model is estimated. This approach provides insights on the various population segments present while evaluating distinct choice regimes for each of these segments. The segmentation approach addresses two concerns: (1) ensures that the parameters are estimated employing the full sample for each segment while using all the population records for model estimation, and (2) provides valuable insights on how the exogenous variables affect segmentation. An Expectation-Maximization algorithm is proposed to address the challenges of estimating the resulting endogenous segmentation based econometric model. A prediction procedure to employ the estimated latent MDCEV models for forecasting is also developed. The proposed model is estimated using data from 2009 National Household Travel Survey (NHTS) for the New York region. The results of the model estimates and prediction exercises illustrate the benefits of employing an endogenous segmentation based MDCEV model. The challenges associated with the estimation of latent MDCEV models are also documented.

Keywords: Multiple discrete continuous models, latent segmentation approaches, daily vehicle type and use decisions, activity type, accompaniment type, and mileage.

1. BACKGROUND

1.1. Multiple-Discreteness

The traditional single discrete choice models are used for examining choice processes where decision makers choose one alternative from the universal choice set of alternatives. However, in a situation where decision makers have the option of choosing several alternatives from the universal set of choice alternatives, the application of single discrete choice model does not represent behavior appropriately. Examples of such multiple discrete choice decision processes include household vehicle type choice, airline carrier choice, grocery item brand choice (such as cookies, cereals, soft drinks, yogurt), and stock selection.

Given the wide range of applications of these multiple-discrete choice processes, it is not surprising that a number of alternative approaches have been proposed to study multiple-discrete choice processes in recent years. One alternative is to employ a single discrete choice model to study these decisions by artificially constructing combination alternatives that consider all possible configurations of the original alternatives. However, as the number of alternatives under consideration increase, the number of *"artificial"* alternatives to be generated increases exponentially (order of 2^K for K alternatives). Another alternative approach often employed is the application of multivariate probit (logit) models that manifest dependency across the various alternatives through correlation in the unobserved component (Manchanda et al., 1999; Edwards and Allenby 2003; Srinivasan and Bhat 2005). A third approach is the one proposed by Hendel and Dube where the multiple-discrete choice process is represented as a series of single discrete choice processes (Hendel 1999; Dube 2004). These three approaches discussed so far examine the multiple discrete problem in the realm of single discrete models i.e. these are clever approaches that extend single discrete structures to study multiple-discrete choice scenarios. These approaches are not only computationally challenging (particularly 1 and 2), but also resort to artificial constructs to model multiple-discrete choice scenarios.

1.2. Kuhn-Tucker Systems

An alternative stream of literature has examined the issue of multiple-discrete choice processes by coupling a continuous component associated with the alternative and a decision maker level budget for the continuous component. These approaches are often referred to as multiple-discrete continuous models. This approach, with its origin in the Kuhn-Tucker (KT) method, was proposed by Wales and Woodland (Wales 1983). These approaches consider a utility function U(x) that is assumed to be quasi-concave, increasing, and continuously differentiable with respect to the continuous component vector x. The observed continuous component vectors are modeled employing a random utility framework while ensuring that the budget constraint is not violated. Given the assumption on U(x), the constraint will actually be binding i.e. continuous component vector is obtained by maximizing the random utility by utilizing the entire continuous component. The KT approach incorporates stochasticity by assuming that U(x) is random and then derives the continuous vector subject to the linear budget constraint by using the KT conditions for constrained optimization. The KT approach constitutes a more theoretically unified and behaviorally consistent framework for dealing with multiple discrete-continuous processes. However, the KT approach did not receive much attention until relatively recently because the random utility distribution assumptions used by Wales and Woodland led to a complicated likelihood function that involves multi-dimensional integration. Bhat introduced a simple and parsimonious econometric approach to handle multiple discreteness based on the generalized variant of the translated constant elasticity of substitution (CES) utility function with a multiplicative log-extreme value error term (Bhat 2005). Bhat's model, labeled the multiple discrete-continuous extreme value (MDCEV) model, is analytically tractable in the probability expressions and is practical even for situations with a large number of discrete consumption alternatives. Since its inception, the MDCEV model has received significant attentions in the transportation community (for a list of studies employing the MDCEV model see Bhat and Eluru 2010).

1.3. Incorporating the Systematic Component

The objective of the current research effort is to contribute to the burgeoning literature on multiple-discrete continuous models by formulating a latent segmentation based MDCEV model that allows for the influence of exogenous variables to vary across the different segments of the population enhancing the heterogeneity captured in the modeling process. An often employed alternative to address the issue of population homogeneity is to consider random components or error correlations in the MDCEV framework (see Pinjari and Bhat 2010; Pinjari 2011). The recent increase in computation power and advances in simulation field have contributed substantially to the use of mixed versions of the MDCEV model (see Munger et al., 2012). However, these approaches focus their attention on the unobserved component of the utility expression. The goal of this paper is to examine an alternative method to address population heterogeneity within the MDCEV model structure.

Prior to enhancing our understanding of the unobserved component, it is necessary to focus our attention on the systematic component (observed variables) of the utility function. A commonly proposed approach to incorporate population heterogeneity is the segmentation of the population into various segments with a segment specific choice model. The natural question that arises is how do we segment the population? The population can be grouped into mutually exclusive segments based on exogenous variables: males and females; individuals with and without access to car; and so on. However, when the analyst is interested in incorporating multiple variables for the segmentation task, the number of segments and segment specific choice models increase the associated computational burden. Further, as the number of mutually exclusive segments increases, the sample size within each segment diminishes rapidly reducing the efficiency in parameter estimation.

An effective solution to the above problem is to consider endogenous segmentation of the population (Bhat 1997). The endogenous segmentation approach allocates decision makers probabilistically to various segments as a function of exogenous variables. Within each endogenously determined segment, a segment specific choice model is estimated. The approach allows us to gather insights on the various population segments present while evaluating distinct choice regimes for each of these segments. The segmentation approach addresses two concerns: (1) ensures that the parameters are estimated employing the full sample for each segment while employing all the population records for model estimation, and (2) provides valuable insights on how the exogenous variables affect segmentation. The approach outlined here forms a subset of latent class models for the multiple-discrete continuous context. There have been a number of studies in the single discrete choice domain in terms of examining latent class models. These latent class models have been applied for unordered systems (Bhat 1997; Greene and Hensher 2003, Anowar et al., 2012) and ordered systems (Eluru et al., 2012). We propose an equivalent latent segmentation approach for the multiple-discrete continuous frameworks in our study.

1.4. Current Research in Context

There have been earlier studies on examining latent class models for multiple-discrete continuous choices. Kuriyama et al. propose a latent segmentation approach for KT systems

(Kuriyama et al., 2010). The study belongs to a stream of research in the environmental economics field (Phaneuf et al., 2000; von Haefen 2003; von Haefen and Phaneuf 2005; Phaneuf and Smith 2005) that has also used the KT approach to study multiple-discreteness. These studies use variants of the linear expenditure system (LES) as proposed by Hanemann and the translated CES for the utility functions, and use multiplicative log-extreme value errors (Hanemann 1978). However, the error specification in the utility function is different from that in Bhat's MDCEV model, resulting in a different form for the likelihood function. The current approach proposes and employs the latent segmentation approach for the MDCEV model. Further, the empirical setting involved in the Kuriyama et al. (2010) study entails estimating generic parameters (i.e. alternative specific parameters are not estimated). This allows for the reduction of the number of parameters estimated in the study – an important criterion in estimating latent class models that are known to exhibit instability in the estimation process. The presence of alternative specific parameters adds to the computational complexity of the estimation process of latent segmentation models (more on this in Section 7). In the MDCEV context also there has been one latent class study (Castro et al., 2011). In this study, the authors consider the latent aspect of choice set generation for individuals. The approach is demonstrated successfully in the context of tour choice and associated mileage. This method, similar to the single discrete approaches for choice set generation, is applicable only in the context where the number of choice alternatives is manageable. In choice scenarios with large number of alternatives, choice set generation based approaches become unmanageable.

To summarize, the proposed study contributes to travel behavior literature in the following ways. The proposed approach is the first implementation of endogenous segmentation for the MDCEV model in extant literature. The model estimation is undertaken using Full Information Maximum Likelihood (FIML) as well as the Expectation Maximization (EM) approach. Second, the latent MDCEV model is applied on the 2009 National Household Travel Survey for the New York region to study non-workers daily decision of vehicle type and usage (represented as miles) in conjunction with activity type and accompaniment choice decisions with a universal choice set of 75 alternatives[1]. Third, the study documents the challenges in the estimation of latent segmentation MDCEV models. Finally, a customized prediction framework for the latent segmentation model that builds on the KT forecasting procedure (see Pinjari and Bhat 2010) is employed for the validating the prediction results for the NHTS dataset.

The reminder of the paper is organized as follows. Section 2 presents the methodology for the endogenous segmentation based MDCEV model; this section also describes the EM approach for estimation and the proposed latent segmentation prediction system. Section 3 provides a brief introduction to the empirical setting. Section 4 presents details on data assembly procedures and sample characteristics. Section 5 presents a contrast between the latent MDCEV model vis-a-vis the traditional MDCEV model. In Section 6, the estimation results of the endogenous segmentation based MDCEV model are presented. The authors document the challenges faced in the estimation of latent segmentation MDCEV model in Section 7. Section 8 provides a discussion of the prediction performance of the traditional MDCEV and the proposed latent MDCEV model. Section 9 summarizes and concludes the paper.

[1] The latent MDCEV model is also estimated on a sample drawn from the 2010 American Time-Use Survey (ATUS) data to study daily activity time-use participation decisions for non-workers. Due to space considerations, only the results from the NHTS dataset are discussed in the paper

2. ECONOMETRIC METHODOLOGY

2.1. Model Structure

Let us consider “S” homogenous segments of the population (1, 2, ..., *S* where S is to be determined). The pattern of decision process within the segment remains identical. However, there are intrinsic differences in the pattern of multiple-discrete continuous choice process across different segments i.e. we have a distinct multiple-discrete continuous choice process for each segment.

2.1.1. Segment specific formulation

Within each segment *s*, we formulate the MDCEV model in its original form (Bhat and Eluru 2010; Bhat 2008). We consider the following functional form for utility in this paper, based on a generalized variant of the translated CES utility function and with the consideration for one outside good (essential Hicksian composite good):

$$U_s(\boldsymbol{x}) = \frac{1}{\alpha_{1s}} \exp(\varepsilon_{1s}) \left\{ (x_1 + \gamma_{1s})^{\alpha_{1s}} \right\} + \sum_{k=2}^{K} \frac{\gamma_{ks}}{\alpha_{ks}} \exp(\psi_{ks} + \varepsilon_{ks}) \left\{ \left(\frac{x_k}{\gamma_{ks}} + 1 \right)^{\alpha_{ks}} - 1 \right\} \qquad (1)$$[2]

where $U_s(\boldsymbol{x})$ is a quasi-concave, increasing, and continuously differentiable function with respect to the consumption quantity (*K*x1)-vector $\boldsymbol{x}$ ($x_k \geq 0$ for all *k* alternatives), and ψ_{ks} ($= \exp(\beta_s' z_{ks})$), γ_{ks} and α_{ks} are parameters associated with alternative *k* in segment *s*. ψ_{ks} represents the baseline marginal utility for segment *s*, z_{ks} represent the vector of exogenous variables in the marginal utility for segment *s*, γ_{ks} enable corner solutions while simultaneously influencing satiation and α_{ks} influences satiation only. Due to the similar role of γ_{ks} and α_{ks} (in terms of allowing for satiation) it is very challenging to identify both γ_{ks} and α_{ks} in empirical applications due to identification challenges (see Bhat 2008 for an elaborate discussion on the issue). Usually, one chooses to estimate satiation using γ_{ks} or α_{ks}.

Depending on the chosen parameter for estimation the alternative utility structures are described as follows:

In the case where only the γ_{ks} parameters are estimated the utility simplifies to

$$U_s(\boldsymbol{x}) = \exp(\varepsilon_{1s}) \ln \left\{ x_1 + \gamma_{1s} \right\} + \sum_{k=2}^{K} \gamma_{ks} \exp(\beta_s' z_{ks} + \varepsilon_{ks}) \ln \left(\frac{x_k}{\gamma_{ks}} + 1 \right) \qquad (2)$$

Similarly, in the case of estimating only α_k the corresponding utility expression collapses to

$$U_s(\boldsymbol{x}) = \frac{1}{\alpha_{1s}} \exp(\varepsilon_{1s}) x_1^{\alpha_{1s}} + \sum_{k=2}^{K} \frac{1}{\alpha_{ks}} \exp(\beta_s' z_{ks} + \varepsilon_{ks}) \left\{ (x_k + 1)^{\alpha_{ks}} - 1 \right\} \qquad (3)$$

[2] In the event of the absence of outside goods in the empirical context, the equation is modified as $U_s(\boldsymbol{x}) = \sum_{k=1}^{K} \gamma_{ks} \exp(\beta_s' z_k + \varepsilon_{ks}) \ln \left(\frac{x_k}{\gamma_{ks}} + 1 \right)$

Following Bhat (Bhat 2005; Bhat 2008), consider an extreme value distribution for ε_{ks} and assume that ε_{ks} is independent of z_{ks} (k = 1, 2, …, K) . The ε_{ks} 's are also assumed to be independently distributed across alternatives with a scale parameter of 1[3]. Let V_{ks} be defined as alternative utility in segment *s*. In that case, the value of V_{ks} according to the two profiles are as follows:

γ -profile

$$V_{ks} = \beta_s' z_{ks} - \ln\left(\frac{x_k^*}{\gamma_{ks}} + 1\right) \; (k \geq 2); \; V_{1s} = -\ln(x_1^* + \gamma_{1s}) \tag{4}$$

α -profile

$$V_{ks} = \beta_s' z_{ks} + (\alpha_{ks} - 1)\ln\left(x_k^* + 1\right); V_{1s} = \left(\alpha_{1s} - 1\right)\ln(x_1^*) \tag{5}$$

Given the V_{ks} values for the two profiles[4], the probability that the individual *q* (*q* = 1, 2, ..., *Q*) has a continuous vector (e_k^*) for the first *M* of the *K* goods (*M* ≥ 1) conditional on the segment choice s is given as follows:

$$P_q(\beta_s, \gamma_s) \,|\, S = P_q\left(e_1^*, e_2^*,\ e_3^*,\ \ldots,\ e_M^*,\ 0,\ 0,\ \ldots,\ 0\right) | \, S = \left[\prod_{i=1}^{M} c_i\right]\left[\sum_{i=1}^{M} \frac{1}{c_i}\right]\left[\frac{\prod_{i=1}^{M} e^{V_{is}}}{\left(\sum_{k=1}^{K} e^{V_{ks}}\right)^{M}}\right](M-1)! \tag{6}$$

It is important to recognize that the individual utility maximization is subject to the binding linear budget constraint that $\sum_{k=1}^{K} e_k^* = E$ where E is the total continuous quantity. The analyst can supply the appropriate V_k values depending on the profile under consideration in the analysis. The proposed analysis approach of the latent segmentation MDCEV will not alter based on the profile employed.

2.1.2. Segment choice formulation

Now we need to determine how to assign the decision makers probabilistically to the segments. The random utility based multinomial logit structure is employed for the segmentation model. The utility for assigning an individual *q* to segment *s* is defined as:

$$W_{qs}^* = \delta_s' y_q + \xi_{qs} \tag{7}$$

y_q is an (*M* x 1) column vector of attributes (including a constant) that influences the propensity of belonging to segment *s*. δ_s is a corresponding (*M* x 1)-column vector of coefficients and ξ_{qs} is an idiosyncratic random error term assumed to be identically and independently Type 1 Extreme

[3] In the presence of price variation across the various alternatives the scale parameter can be identified. However, in the absence of price variation the scale parameter is unidentifiable and is set to 1 for convenience (see Bhat 2008 for extensive discussion)

[4] In our empirical context we found that the MDCEV model based on the γ -profile offered substantially better fit to compared to the MDCEV model with the α -profile.

Value distributed across individuals *q* and segment *s*. Then the probability that individual *q* belongs to segment *s* is given as:

$$P_{qs} = \frac{\exp(\delta_s^{'} y_q)}{\sum_{k=1}^{S} \exp(\delta_k^{'} y_q)} \tag{8}$$

Based on the above discussion, the unconditional probability of multiple-discrete continuous choice pattern:

$$P_q = \sum_{s=1}^{S} \left[\left(P_q \left(e_1^*, e_2^*, e_3^*, \ldots, e_M^*, 0, 0, \ldots, 0 \right) | S \right) * P_{qs} \right] \tag{9}$$

The log-likelihood function for the entire dataset is provided below:

$$L = \sum_{q=1}^{Q} P_q \tag{10}$$

The parameters to be estimated in the model are β (composed of $(\beta_1, \beta_2, \ldots \beta_S)$), γ (composed of $(\gamma_1, \gamma_2, \ldots \gamma_S)$) or α (composed of $(\alpha_1, \alpha_2, \ldots \alpha_S)$) and δ (composed of ($\delta_1, \delta_2, \ldots \delta_S$)) for each *s* and the number of segments *S*[5].

The model estimation approach begins with a model considering two segments. The final number of segments is determined by adding one segment at a time until further addition does not enhance intuitive interpretation and data fit. The data fit is measured using (1) Bayesian Information Criterion (BIC), (2) Akaike information criterion (AIC) and (3) Akaike information criterion corrected (AICc).

2.2. Model Estimation

The estimation of latent class models using quasi-Newton routines can be computationally unstable (Bhat 1997). A commonly employed approach to address the challenges involved in optimization of the log-likelihood function for latent class models is the EM algorithm. EM algorithm employs an iterative approach consisting of two steps: Expectation (E) step and Maximization (M) step. In the E step the segment allocation variables (δ) are estimated based on the observed data and in the M step current iteration parameters are updated by maximizing the likelihood employing the segment allocation variables (δ) estimated in the E step (Bhat 1997; Kuriyama et al., 2010). The EM algorithm is employed as follows:

(1) Starting values for β, γ and δ are assumed; based on the assumption the segment membership function is computed in the Bayesian fashion as $\tilde{P}_{qs} = \frac{\left[\left(P_q(\beta_s, \gamma_s) | S \right) * P_{qs} \right]}{\sum_{s=1}^{S} \left[\left(P_q(\beta_s, \gamma_s) | S \right) * P_{qs} \right]}$ (11)

(2) δ vector is updated while maintaining the β, and γ vectors to remain constant by maximizing a slightly modified version of the log-likelihood function $\sum_{q=1}^{Q} \sum_{s=1}^{S} \tilde{P}_{qs} * P_{qs}$ (12)

[5] To be sure $\beta_s, \gamma_s, \alpha_s$ represent all the K elements of corresponding segment specific vectors. The subscript k is suppressed for ease of notation.

(3) Employing the updated δ vector, new segment membership values ($\tilde{P}_{qs}$) are computed.

(4) β , and γ vectors are updated while maintaining the δ vector to remain constant by maximizing the function $\sum_{q=1}^{Q} \sum_{s=1}^{S} \tilde{P}_{qs} * \left(P_q(\beta_s, \gamma_s) \mid S \right)$ (13)

(5) The procedure is repeated until there is no significant difference in the parameters or the log-likelihood function.

The procedure does not provide the standard error of the parameter estimates. To generate standard errors of the estimates, we provide the convergence solution from the EM approach as initial values and run the Full Information Maximum Likelihood (FIML) model. In our experience, we found the EM approach to be very slow. Hence, we used it to generate the very initial specification for the latent MDCEV model. After we achieved a reasonable set of starting attributes for a stable latent MDCEV we shift to FIML model estimation procedure which was substantially faster than the EM approach. The EM and FIML routines for latent MDCEV models were programmed in Gauss.

2.3. Model Prediction

We also outline a prediction framework for the proposed latent MDCEV model. The prediction process builds on the prediction framework developed for the MDCEV model (Pinjari and Bhat 2010). Specifically, the following approach is employed to predict consumption patterns:

(1) Generate consumption patterns for individuals by employing the segment specific MDCEV models (involves the influence of random component – so repeat K times)

(2) Generate the probability measure for the individual segment membership using the latent segmentation model.

(3) Assign individuals to segment based on their segment membership probabilities by drawing uniform random numbers.

(4) Allocate the consumption patterns to individuals depending on their segment choice in step 3 and consumption patterns obtained in step 1.

(5) Repeat the process (Step 1 to 4) multiple times (L) and compute the average and standard deviation of the resulting consumption patterns to generate a range on the predicted participation. We examine the influence of various values of (K*L).

3. EMPIRICAL SETTING

In our research effort, we focus our attention on short-term vehicle fleet allocation decisions. Specifically, we examine the role of activity type and accompaniment type on vehicle type and usage decisions. The NHTS 2009 data indicates that the vehicle occupancy levels for shopping and social/recreational activities are 1.78 and 2.20 respectively indicating the inherent tendency among individuals to pursue these activities with a companion. Moreover, vehicle miles of travel for social/recreational activities, family and personal errands and other activities are 10.9, 10.6 and 5.4, respectively; implying that activity type affects mileage decisions (Santos et al., 2011). Earlier literature has also found evidence toward increased likelihood of engaging larger vehicles (like SUV or Van) when multiple passengers are engaged (Paleti et al., 2012). In summary, it is plausible to consider strong interactions the following choice dimensions - vehicle type, activity type, accompaniment type and usage.

To be sure, several studies have examined a subset of dimensions identified above. A set of studies have focussed on daily activity participation decisions including activity type and accompaniment type decisions (for example, see Kapur and Bhat 2007; Carrasco and Miller 2009; Ferdous et al., 2010). Only recently research studies exploring short-term vehicle type choice, accompaniment type and tour length decisions have been undertaken (see Paleti et al., 2012*;* Konduri et al., 2011). These studies focus on tour as an entity for participation while neglecting the inherent activities pursued. Further, only Paleti et al. (2012) has examined mileage choices in the context of accompaniment type. The study employs a system of simultaneous equations to generate the correlation across the various dimensions including tour complexity, passenger accompaniment, vehicle type and tour length. The approach, while simulation free, still resorts to coupling of choices through the unobserved component.

In our current study, we propose a unified model that simultaneously allows for competition across the various alternatives within a random utility based approach while considering the daily vehicle type and usage decisions for every activity type and accompaniment type combination. Towards this end we focus on three dimensions: (1) vehicle type, (2) activity type and (3) accompaniment type. To consider vehicle type, we recast the vehicle type choice process as a travel mode choice process by considering the various travel mode alternatives (transit, walking/bicycling) and replacing the private vehicle alternative with various vehicle type options that are available to individuals. We recognize that the available private vehicle alternatives are dependent on the household vehicle ownership decisions. The activity type and accompaniment decisions are directly obtained from NHTS data responses. Thus the three dimensions: (1) travel mode that implicitly considers vehicle type, (2) activity purpose and (3) accompaniment type are jointly analyzed by generating combination alternatives (an example alternative: SUV- shopping- with household members). The continuous component essential for the MDCEV budget constraint is considered through the mileage dimension for each discrete alternative combination.

4. DATA SOURCE AND SAMPLE FORMATION

The data for our research effort is drawn from National Household Travel Survey (NHTS) data conducted in 2008-2009 for New York, Northern New Jersey and Long Island region. The survey compiled information on individual and household socio-demographics, residential location characteristics and daily travel attributes including out-of-home activity episode type, the day and month on which the activity is undertaken, travel mode for every episode (including vehicle type information for automobile users) and accompanying person information (alone, household or non-household members) for the episode.

For the purpose of our analysis we restrict our attention to non-work activity purposes classified into five main categories: (1) Shopping, (2) Social and recreational, (3) Transporting someone, (4) Meals and (5) Others. The *travel mode alternatives* are characterized as: (1) Public transit, (2) Walk/bike (these two modes are available for everyone) and three privately owned vehicle types: (3) Car, (4) SUV and (5) Other vehicles (including Van and pick up). The vehicle type dimensions are appropriately matched with the household vehicle ownership information (*i.e.* if a household does not own a SUV, the individual will not have alternatives corresponding to SUV available to him/her). The *accompaniment dimension* is classified as: (1) Alone, (2) With household member and (3) With household members and non-household members. Overall, these categories result in 75 discrete alternatives (5*5*3). The mileage component associated with these discrete alternatives is provided as the continuous component of the MDCEV model.

The sample formation exercise involved a series of transformations on the original NHTS travel data set. First, the respondents that participated in the work activity were selected. Second, for the worker sample, information on activity purpose, travel mode and accompaniment type were gathered for out-of-home activities on weekdays. Subsequently, the combination alternatives across the three dimensions and the corresponding mileage metrics were generated. Third, the respondent related socio-demographics, residential location and contextual characteristics (day of week and the season of travel day) were appropriately appended to the database. Fourth, consistency checks were performed on the sample, and records with missing or inconsistent data were eliminated. Finally, a small hold-out sample was created to allow for validation comparison of the proposed model.

5. MODEL EVALUATION

5.1. Model Fit

The model estimation of the latent MDCEV structure began with an estimation of the traditional MDCEV model. Subsequently a latent segmentation model with two segments was estimated. Then, we continued adding additional segments to the model as long as there was the additional segment provided an improvement in the overall log-likelihood function. In this process, we estimated four model structures: (1) MDCEV model, (2) Latent MDCEV model with two segments (latent MDCEV 2), Latent MDCEV model with three segments (latent MDCEV 3), and Latent MDCEV model with four segments (latent MDCEV 4). Since the various models are not nested within one another, the Bayesian Information Criterion (BIC), Akaike information criterion (AIC) and Akaike information criterion corrected (AICc) are employed to compare model performance (see Schwarz 1978; Akaike 1977; Burnham and Anderson 2004)[6]. The log-likelihood, BIC, AIC and AICc measures along with the parameter set size for the four model systems are presented in Table 1. The model fit measures presented clearly highlight that the three segment model outperforms the other models substantially. It is also important to note that all the latent segmentation models significantly outperform the traditional MDCEV model. These model fit measures substantiate our hypothesis that relaxing the population homogeneity assumption enhances the statistical fit of the data. For the sake of brevity, the discussion of model results is confined to the three segment latent MDCEV model.

The reader should note that the model specification was arrived at through a systematic process of removing statistically insignificant variables and combining variables when their effects were not significantly different. It was found that the dummy representation of continuous variables offered superior fit compared to the corresponding linear variables.

5.2. Latent MDCEV Framework versus Exogenously Segmented MDCEV Framework

The latent MDCEV model allows us to incorporate population heterogeneity through endogenous segmentation. Based on our discussion in section 5.1, it is evident that the three segment latent MDCEV model statistically outperforms the MDCEV model. To further highlight the advantages of the latent MDCEV model we compare its performance with exogenous segmentation based MDCEV model. Specifically, we split the dataset into distinct sub-datasets based on exogenous variables. Subsequently, we estimate MDCEV models for each of the sub-datasets and compare the model fit with the three segment MDCEV model. The exogenous variables considered for segmentation include: (1) Males and Females, (2) Age 21 and under and Age over 21 years, (3) Household size less than 3 and household size 3 and above, (4)

[6] The BIC for a given empirical model is equal to − 2ln(L) + K ln(Q), where ln(L) is the log-likelihood value at convergence, K is the number of parameters, and Q is the number of observations. AIC is represented by 2K - 2ln(L) and AICc is defined as AIC + 2K (K+1)/(Q – K – 1).

Residential density less than 10,000 per square mile and residential density above 10,000 per square mile, (5) Gender and Age combinations (from 1 and 2) and (6) Gender and household size combination (from 1 and 3). The results for the log-likelihood values for the six variable combinations are presented in Table 2. The comparable log-likelihood to the three segment latent MDCEV model can be computed by just summing up the different sample log-likelihoods. The comparison of these log-likelihood measures with the latent segment MDCEV model (-16283.86) clearly shows the efficacy of the endogenous segmentation approach relative to the exogenous approach. The approach to undertake the exogenous segmentation – though easier to achieve - is bound to be inefficient relative to the endogenous approach.

5.3. Segmentation Properties of the Latent MDCEV Three framework

The three segment MDCEV model estimations can be used to generate information regarding the aggregate percentage population share across the three segments based on the segment membership component. For the latent MDCEV 3 model we observe the following membership shares: Segment one – 56.4%, Segment two – 22.0% and Segment three – 21.6%. The population membership shares highlight the significantly heterogeneous nature of the population sample and require a careful consideration for policy analysis.

To provide further insight on the distinct profiles of the segments we can also determine the mean values of the segmentation variables in the three segments. To compute these measures we will employ the following notation (see Bhat 1997; Anowar et al., 2012 for similar computation for the single discrete context):

$$\tilde{y}_s = \frac{\sum_q P_{qs} y_q}{\sum_q P_{qs}} \qquad (14)$$

Where $\tilde{y}_s$ represents the mean segmentation value of the segmentation variables y. The computation of these measures is based on the segmentation component parameters of the MDCEV 3 model presented in Table 3. The $\tilde{y}_s$ measures computed for all exogenous variables affecting segmentation are presented in Table 4. The variable shares across the different segments offer intuitive distributions for the various segmentation variables. For example, male variable (positive effect for segment three) indicates a slightly higher proportion of males allocated to segment three compared to the population measure; at the same time the other two segments have a slightly smaller share of males compared to the population share. On the other hand, the spring variable (negative coefficient for segment two) reduces the overall likelihood of spring day being allocated to segment two, while increasing the likelihood for allocation to either segment one or three. The impact for spring variable is of a higher magnitude than that for the male variable because of the larger magnitude (-0.86 versus 0.34).

6. MODEL RESULT DISCUSSION

The model results for the three-segment MDCEV consists of four components: (1) latent segmentation component (Table 4), (2) segment one mileage profile (Table 5A), (3) segment two mileage profile (Table 5B) and (4) segment three mileage profile (Table 5C).[7]. In the model specification, several types of variables were considered including: (1) individual demographics (gender, age, race and education level), (2) household demographics (household size, presence of children and family income), (3) household location variables (urban areas and residential density) and (4) contextual variables (day of the week and seasons). In the results presentation for MDCEV components every row represents an activity purpose, travel mode and accompany type dimension; while every column represents the variable effects on the alternatives. Further, a '-' entry indicates the absence of a significant effect of the variable on the corresponding mileage usage utility.

6.1. Segmentation Propensity Component

The latent segmentation model plays the role of assigning individuals probabilistically to the various segments in the latent MDCEV model. In our segmentation component, the utility corresponding to the first segment is chosen to be the base. The segment membership is influenced by individual demographics (age and gender), household demographics (household size and residential density), and contextual variables (spring).

The results indicate that the second segment membership is positively influenced by individuals under 21 years, and larger family size. The membership is negatively influenced by the season dummy variable corresponding to spring season. On the other hand, males are more likely to be assigned as the member of the third segment while individuals residing in dense neighborhoods are less likely to be assigned to the third segment. The endogenous segmentation model presented here allows us to efficiently generate three segments based on 5 exogenous variables. An exogenous segmentation might have required us to segment the population into at least 32 segments while estimating distinct MDCEV models for each segment.

6.2. Segment One Mileage Profile

Activity travel profile for the various segments of the MDCEV 3 model are influenced by individual and household demographics, and contextual variables.

6.2.1. Individual demographics

The individual demographic variables influencing the segment one activity travel profile include gender, race, education and age. The gender variable impact indicates that male non-workers have a higher tendency to choose an SUV or Van as their private vehicle alternative (see Mohammadian and Miller 2003; Paleti et al., 2012 for similar results). The impact of race variable indicates that Caucasian individuals are more likely to pursue out-of-home meals and social/recreation activities compared to individuals of other race. Non-workers with university education are more likely to participate in transport someone while at the same time are less likely to pursue activities with household members.

The age-related variables have a significant association with non-worker activity travel participation profile. The results indicate that individuals older than 22 years are less likely to travel for social/recreational activity, meals, and other activities compared to individuals 21 and younger. Further, these individuals are less likely to pursue activities with non-household members. The results are along expected lines because these individuals are more likely to have

[7] The constant and the gamma parameters for the MDCEV model are not presented due to space considerations. These parameter tables are available with authors. The constants and gamma parameters were also estimated along the various dimensions (as opposed to estimating 2*75 parameters for each segment)

familial responsibilities compared to the younger individuals. Older individuals (aged more than 60 years) are also less likely to be accompanied with non-household members. The result indicates that older individuals prefer to be more involved in activities with their family. On the other hand, these people are less probable to choose walk/bike or public transit as their travel mode respectively. Given their physical condition, it is intuitive that the older individuals are disinclined to employ non-motorized and public transit modes.

6.2.2. Household demographics

Among the household socio-demographic variables, household size, presence of children, household income, residential location and residential density affect the activity travel process. The impact of household size offers interesting insights. Individuals from larger households are more likely to pursue the transport someone alternative and employ either the Van alternative or non-private vehicle alternatives for travel (see Eluru et al., 2010 for similar results). The results indicate two subtle patterns of travel for activity participation. Individuals who can afford vehicle ownership are likely to use larger vehicles (Van) while those individuals that cannot afford multiple vehicles are likely to rely on the non-auto travel modes.

The next household demographic attribute examined is the presence of a child in the household. The variable is introduced as four dummy indicators: (1) household with children less than 5 years old, (2) household with children between 6 and 15 years, (3) household with children between 16 and 21 years and (4) household with no children. As you would expect, presence of children increases the likelihood of transport someone activity participation. Typically, adults are responsible for chauffeuring of children to/from school and other non-school activities. The results for the presence of children variables are consistent with this assumption (similar to Paleti et al., 2012). In terms of accompaniment type, presence of children increases activity participation with household members. In terms of travel mode, households with children aged between 6 to15 years are more likely to choose Van for travel. It is reasonable that households with young children choose larger vehicles (see Eluru et al., 2010; Paleti et al., 2011; Cao et al., 2006 for similar results).

Household income variables indicate that individuals from households with more than $40,000 income are less likely to choose public transit as their daily travel mode; while the ones with high income (more than $70,000 per year) are more likely to be involved in social/recreational activities. Household location characterized as urban area, and residential density greater than 10,000 residents per square mile were introduced in the baseline utility to study the impact of land-use on activity travel process. Non-workers in segment one who live in urban or high residential density areas are less likely to participate in a transport someone activity. Due to enhanced connectivity, it is possible that individuals are less likely to be pursuing transport someone relative to individuals residing in suburban regions. On the other hand, more frequent, accessible public transport services and proximity to activity centers in high density neighborhoods make it more practical to use transit and/or non-motorized modes.

6.2.3. Contextual Variables

In segment one, none of the contextual variables had a statistically significant effect on the activity travel patterns.

6.3. Segment Two Mileage Profile

For the sake of brevity, only major differences in exogenous variable effects are discussed for segments two and three.

6.3.1. Individual demographics

In segment two, men are less likely to travel with an SUV relative to other modes. In this segment, it appears that female members of the family are assigned the responsibility of transporting others. The results for the age-related indicators show that people aged more than 22 years are more likely to pursue transport someone activity. This is presumably due to the possibility that at this age they are more responsible compared to the younger people, and might have their family members who need to be transported. We also observe that individuals between the years 22 through 60 are more inclined to pursue activities alone in segment two. Another interesting impact of the age variable is the disinclination of older individuals in employing the SUV vehicle for non-work activities.

6.3.2. Household demographics

In the second segment the coefficient of household size variable indicates that with increasing household size the probability of transporting someone decreases (different from segment one). Segment two is composed of younger individuals who are unlikely to participate in transport someone activity. Further, as household size increases, the propensity of accompanying non-household member in a daily trip increases (see Paleti et al., 2012 for similar results). The presence of younger individuals encourages activity participation with non-household members. This is further substantiated by the impact of presence of children aged less than 5 years in the household. Individuals with children pursue activities with household members. Similar to segment one, presence of at least one child between 6 to 15 years results in increased participating in transport someone activity purpose.

6.3.3. Contextual Variables

Non-workers in the second segment during summer have a higher tendency to travel for social/recreational purposes whereas in spring and winter they are more involved in transport someone activity. It is interesting to note that during spring individuals are less likely to pursue activities with household members. The exact reasons for these impacts require additional investigation. During winter individuals prefer to use SUV compared to other months probably because larger vehicles offer better control and improve the driving experience in winter. The weekday dummy variable for Friday indicates a disinclination to pursue transport someone activity on Fridays. Further, the results indicate that on Fridays SUVs are less likely to be employed. It is possible that on Fridays individuals travel to activity centers in denser neighborhoods where parking concerns might make SUVs less preferable.

6.4. Segment Three Mileage Profile

6.4.1. Individual demographics

Male non-workers in the third segment are less likely to pursue activities with non-household members. People with higher education level are more likely to be accompanied with their household members. It is possible that these individuals are likely to be in households with higher employment and busier schedules thus reducing the likelihood of activity participation alone or with non-household members.

6.4.2. Household demographics

In segment three, as household size increases, propensity for transport someone increases while propensity for Van usage reduces. Presence of a child is one of the variables at the household level which has significant effect on mileage usage. Individuals with at least one child aged less than 5 years old are more likely to pursue activities with household members. Further, existence of a child between 6 to 15 years increases the tendency of choosing a Van or walk/bike as a travel alternative in this segment.

The next household socio-demographic attribute considered is the effect of household annual income. The coefficient of this variable indicates that households with medium income (annual income between $40,000 and $70,000) are less likely to participate in social/recreational activities. Also, these individuals are less inclined to walk or bike. Further, as you would expect, individuals with annual income more than 70,000 are less likely to choose public transport as their travel mode. The residential location variable indicates that individuals from urban areas are less inclined to accompany their household members.

6.4.3. Contextual Variables

During spring, people are more likely to travel with their household members and opt for Van as their transportation mode while they are likely to prefer SUVs on Fridays. The impact discussed here is in contrast to the impact observed in segment two for the Friday variable. The potential for such distinct variable effects across different segments provides further evidence to the presence of significant population heterogeneity.

Overall, we see that the three segments exhibit distinct activity travel profiles substantiating the hypothesis that distinct activity travel profiles exist in the population. The conventional approach that restricts the activity travel profile to be the same across the population ignores this potential variation and arrives at exogenous variable impacts that are inaccurate.

7. ESTIMATION CHALLENGES

The estimation of latent segmentation models poses significant challenges in pinning down the influence of exogenous variables on the segmentation (MNL model) and within segment choice (MDCEV) components. In this section, we document the challenges we faced in the estimation of the latent MDCEV model. The main challenge arises from the issue of empirical identification. Based on the problem formulation, theoretically, the analyst should be able to estimate the impact of a particular variable in the latent segmentation model as well as the segment specific MDCEV models (while accounting for base variables appropriately). However, the estimation of all theoretically plausible impacts are not always possible due to empirical identification issues i.e. the data does not support their estimation because these parameters are not different from 0. Consider the impact of male variable in the three segment latent MDCEV model. The influence of the male variable can be estimated in the segmentation model as well as the segment specific models. A priori, we have no information to suggest that the variable can be restricted to either segmentation or segment specific models. Hence, theoretically, we should be able to estimate two effects of the influence of male variable in the segmentation component – segment two and segment three (segment one is base) and 10 effects of the male variable per segment in the MDCEV segment level models[8]. So the total number of effects that we could potentially estimate would amount to 32 (2 + 3 * 10). We expect a large portion of these 32

[8] The number of variable effects is limited to 10 as we focus on the impact of variable on each dimension rather than the actual combination alternatives. The approach allows for parsimonious specification structure and is widely used in choice processes with large number of alternatives (for example Kapur and Bhat 2007; Eluru et al., 2010).

parameters to not have a statistically significant impact. The typical approach to employ in the traditional MDCEV model would be to try to estimate the 32 variables and discard variables that are insignificant. However, in the latent MDCEV model, adding 32 variables simultaneously would lead to empirical identification resulting in lack of convergence of the log-likelihood optimization routine. In the absence of a converged solution it is not possible to estimate the standard errors for the parameters. This does not happen in the MDCEV model because of the well-defined nature of the log-likelihood function. The latent MDCEV model (or for that matter all latent segmentation models) are weighted averages of multiple log-likelihood functions. Hence, these functions are not as well behaved as their non-latent counterparts.

In fact, the challenges with the latent MDCEV are similar to the empirical identification issues observed in the estimation of simulated maximum likelihood (Cherchi and Guevara 2012). In the simulated maximum likelihood optimization routine it is very likely that the analyst finds optimization routine convergence issues because of the flatness of the log-likelihood function. In the latent case, we observe that the log-likelihood function at the initial stages of the latent segmentation model is relatively flat thus making it hard to identify the impact of exogenous variables. It is comforting to note that once, we have established a convergent set of parameter estimates, it is easier to build on the specification and identify factors that influence the various components. The EM method proposed is useful in this regard, particularly for the initial specification set up.

In our analysis, to address the aforementioned issues we followed the following guidelines for model estimation. First, we started the estimation with a stable MDCEV model to identify the exogenous variables that are likely to influence the choice process. Second, we estimated the latent MDCEV model with two segments by starting with distinct exogenous variables in the segmentation component and the segment specific components. Thus, we minimized the potential empirical identification problem. Third, once we obtained a stable latent MDCEV two segment model, we added one variable at a time to complete the specification process. The process was repeated for the three segment and four segment models. It is important to note that even when the log-likelihood function does not converge the parameter values when the iterations stop provide useful information on the plausible parameter values[9].

Overall, from our experience, the task of formulating the MDCEV model to incorporate systematic heterogeneity through latent segmentation is less challenging compared to the task of empirically estimating the latent segmentation MDCEV model due to a host of empirical identification issues arising when the number of alternatives and parameters sought be estimated are large. The algorithm routines have been coded in GAUSS for our analysis.

8. MODEL VALIDATION

As discussed in section 5.1, the log-likelihood measures and model estimates clearly highlight the superior data fit offered by the three segment MDCEV model. To examine the performance of the latent models in prediction, we undertake a comprehensive validation exercise. We implement the prediction framework discussed in section 2.3. For every individual in the dataset, we generate participation and mileage values were examined for various K*L values (2500, 5000, 25000 and 50000). Beyond the value of 5000, there was little to no change in the mean

[9] For instance, if a parameter estimate has a value of 0.001 around the 200th iteration it is most likely going to have an insignificant effect. This is a useful guideline (particularly for dummy variables). However, these guidelines might not be applicable for continuous variables with large range (such as land use mix).

predicted values while the standard deviations reduced for larger K*L values. In fact, for the 50000 value, the standard deviations were within 0.05% of the average values. Hence, we can ascertain that the means are computed with high level of accuracy. For the sake of brevity we limit ourselves to presentation of the results for 50000 repetitions. The validation exercise was conducted on two data samples: (1) estimation sample from 2009 NHTS data (1937 observations), and (2) hold out sample from 2009 NHTS data (378 observations).

Tables 6 and 7 present the comparison for the MDCEV model, latent MDCEV 2 segment and latent MDCEV 3 segment models for the 2009 NHTS data. The prediction exercise generates outputs for participation mileage values for the 75 alternatives. However, to undertake a comparison in a meaningful way, the participation measures are aggregated across the various accompaniment types, activity purpose and travel mode dimensions. The predicted participation outcomes are compared with the actual observed participation in the estimation sample (Table 6) and hold out sample (Table 7). To compute an overall metric of error in prediction the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for each model are calculated. The results clearly highlight the improved accuracy offered by the latent segmentation models. The lowest values of RMSE and MAE measures are obtained for the 3 segment latent MDCEV model. The two-segment MDCEV model also provides an improved data fit compared to the traditional MDCEV model. It is encouraging to note that even in the hold out sample, a similar trend is observed. Further, it is interesting to note that the error bands in prediction for activity participation across regimes are satisfactory considering the size of the universal choice set (75). Overall, the results illustrate that the latent MDCEV model offers improved prediction capabilities relative to the traditional MDCEV model.

9. SUMMARY

The objective of the current research effort is to contribute to the burgeoning literature on multiple-discrete continuous models by formulating a latent segmentation based MDCEV model. The MDCEV model in its traditional form restricts the exogenous parameter effects across the population i.e. there is an implicit population homogeneity assumption within the model structure. In the event that this assumption is violated the MDCEV model parameter estimates are likely to be biased. An effective approach to incorporate population heterogeneity is to consider endogenous segmentation of the population. The endogenous segmentation approach allocates decision makers probabilistically to various segments as a function of exogenous variables. Within each endogenously determined segment, a segment specific choice model is estimated. The segmentation approach ensures that the parameters are estimated employing the full sample for each segment while employing all the population records for model estimation, and provides valuable insights on how the exogenous variables affect segmentation. The proposed approach is the first implementation of endogenous segmentation for the MDCEV model in extant literature. The model estimation is undertaken using Full Information Maximum Likelihood (FIML) as well as the Expectation Maximization (EM) approach.

The proposed latent MDCEV model is applied to data drawn from the 2009 National Household Travel Survey for the New York region. In our empirical context, the latent segmentation based MDCEV model estimation process involved estimating four model structures: (1) MDCEV model, (2) Latent MDCEV model with two segments, (3) Latent MDCEV model with three segments and (4) Latent MDCEV model with four segments. The MDCEV model with three segments offered the superior fit based on a host of measures. In the model specification, several types of variables were identified to influence choice process such as: (1) individual demographics (gender, age, race and education level), (2) household demographics (household size, presence of children and family income), (3) household location variables (urban areas and residential density) and (4) contextual variables (day of the week and seasons). The model estimation results highlight how the latent MDCEV model allows exogenous variables to exhibit distinct activity participation profiles across various segments. This is clearly illustrated by the varying coefficients for the same exogenous variable the different segments (see for instance, male or income) across.

Further, to examine the performance of the latent models in prediction, a prediction framework for latent segmentation based models was proposed. Through the prediction framework, we undertake a comprehensive validation exercise on two datasets: (1) estimation sample and (2) validation sample. For the 2009 NHTS sample to undertake a comparison in a meaningful way, the participation measures are aggregated across the various accompaniment types, activity purpose and travel mode dimensions. The predicted participation outcomes are compared with the actual observed participation in the estimation sample and hold out sample. To compute an overall metric of error in prediction the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for each model were calculated. The estimation and validation results highlight the importance of incorporating population heterogeneity in the modeling framework within the MDCEV model structure. The latent MDCEV models offer improved data fit as well as improved predictive capabilities. The study also documents the challenges with estimation of latent MDCEV models – a useful exercise for transportation modellers estimating latent segmentation models of various kinds.

The current study is not without limitations. The continuous budget constraint in the MDCEV model is exogenous and assumes that the overall mileage component is "known" to the analyst. The assumption is quite restrictive and in order to enhance our understanding of the choice process it will be useful to endogenize the budget information.

ACKNOWLEDGEMENTS

The corresponding author would like to acknowledge financial support from Natural Sciences and Engineering Research Council (NSERC) of Canada under the Discovery Grants program for undertaking the research. The authors are also grateful to Prof. Chandra Bhat and Dr. Abdul Pinjari whose MDCEV related codes were altered for use in our paper. The authors would also like to acknowledge valuable feedback from three anonymous reviewers on a previous version of the paper.

REFERENCES

Akaike, H., 1977. On entropy maximization principle. In: Krishnaiah, P.R. (Eds.). Applications of Statistics, North-Holland, Amsterdam, pp. 27–41.

Anowar, S., Yasmin, S., Eluru, N., and Miranda-Moreno L., 2012. Analyzing Car Ownership in Two Quebec Metropolitan Regions: A Comparison of Latent Ordered and Unordered Response Models. Technical Paper, Department of Civil Engineering and Applied Mechanics, McGill University.

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., 2005. A Multiple Discrete-Continuous Extreme Value Model: Formulation and Application to Discretionary Time-Use Decisions. Transportation Research Part B 39 (8), 679-707.

Bhat, C.R., 2008. The Multiple Discrete-Continuous Extreme Value (MDCEV) Model: Role of Utility Function Parameters, Identification Considerations, and Model Extensions. Transportation Research Part B 42 (3), 274-303.

Bhat, C.R., Eluru, N., 2009. A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling. Transportation Research Part B 43 (7), 749-765.

Bhat, C.R., Eluru, N., 2010. The Multiple Discrete-Continuous Extreme Value (MDCEV) Model: Formulation and Applications. Choice Modelling: The State-of-the-Art and the State-of-Practice. In: Proceedings from the inaugural International Choice Modelling Conference, pp. 71-100.

Bhat, C.R., Sen, S., Eluru N., 2009. The Impact of Demographics, Built Environment Attributes, Vehicle Characteristics, and Gasoline Prices on Household Vehicle Holdings and Use. Transportation Research Part B 43 (1), 1-18.

Burnham, K.P., Anderson, D.R., 2004. Multi-model Inference: Understanding AIC and BIC in Model Selection. Sociological Methods and Research 33, 261–304.

Cao, X., Mokhtarian, P.L., Handy, S.L., 2006. Neighborhood Design and Vehicle Type Choice: Evidence from Northern California. Transportation Research Part D 11 (2), 133-145.

Carrasco, J.A., Miller, E.J., 2009. The Social Dimension in Action: A Multilevel, Personal Networks Model of Social Activity Frequency between Individuals. Transportation Research Part A 43 (1), 90-104.

Castro, M., Eluru, N., Bhat, C.R., Pendyala R.M., 2011. A Joint Model of Participation in Non-Work Activities and Time-of-Day Choice Set Formation for Workers. Transportation Research Record 2254, 140-150.

Cherchi, E., Guevara, C.A., 2012. A Monte Carlo Experiment to Analyze the Curse of Dimensionality in Estimating Random Coefficients Models with a Full Variance–covariance Matrix. Transportation Research Part B 46 (2), 321-332.

Dube, J.P., 2004. Multiple Discreteness and Product Differentiation: Strategy and Demand for Carbonated Soft Drinks. Marketing Science 23 (1), 66–81.

Edwards, Y.D., Allenby, G.M., 2003. Multivariate Analysis of Multiple Response Data. Marketing Research 40, 321-334.

Eluru, N., Bagheri, M., Miranda-Moreno, L., Fu, L., 2012. A Latent Class Modelling Approach for Identifying Vehicle Driver Injury Severity Factors At Highway-Railway Crossings. Accident Analysis & Prevention 47 (1), 119-127.

Eluru, N., Bhat, C.R., Pendyala, R.M., Konduri, K.C., 2010. A Joint Flexible Econometric Model System of Household Residential Location and Vehicle Fleet Composition/Usage Choices. Transportation: TRB Special Issue 37 (4), 603-62.

EPA, 2009. U.S. Greenhouse Gas Inventory Report. US Environmental Protection Agency.

Ferdous, N., Eluru, N., Bhat, C.R., Meloni, I., 2010. A Multivariate Ordered Response Model System for Adults’ Weekday Activity Episode Generation by Activity Purpose and Social Context. Transportation Research Part B 44 (8-9), 922-943.

Greene, W.H., Hensher, D.A., 2003. A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit. Transportation Research Part B 37 (8), 681-698.

Hanemann, W.M., 1978. A Methodological and Empirical Study of the Recreation Benefits from Water Quality Improvement. Ph.D. Dissertation, Department of Economics, Harvard University.

Hendel, I., 1999. Estimating Multiple-Discrete Choice Models: An Application to Computerization Returns. Review of Economic Studies 66 (2), 423–446.

Kapur, A., Bhat, C.R., 2007. On Modeling Adults' Daily Time Use by Activity Purpose and Accompaniment Arrangement. Transportation Research Record 2021, 18-27.

Konduri, K.C., Ye, X., Sana, B., Pendyala, R.M., 2011. A Joint Tour-Based Model of Vehicle Type Choice and Tour Length. Transportation Research Record: Journal of the Transportation Research Board, 28-37.

Kuriyama, K., Hanemann, W.M., Hilger, J.R., 2010. A latent Segmentation Approach to a Kuhn–Tucker Model: An Application to Recreation Demand. Environmental Economics and Management 60 (3), 209-220.

Manchanda, P., Ansari,A., Gupta, S., 1999. The Shopping Basket: A Model for Multi-Category Purchase Incidence Decisions. Marketing Science 18, 95-114.

Mohammadian, A., Miller, E.J., 2003. Empirical Investigation of Household Vehicle Type Decisions. Transportation Research Record 1854, 99–106.

Munger, D., L'ecuyer, P., Bastin, F., Cirillo, C., Tuffin, B., 2012. Estimation of the Mixed Logit Likelihood Function by Randomized Quasi-monte Carlo. Transportation Research Part B 46 (2), 305-320.

Paleti, R., Eluru, N., Bhat, C.R., Pendyala, R.M., Adler, T.J., Goulias, K.G., 2011. Design of Comprehensive Microsimulator of Household Vehicle Fleet Composition, Utilization, and Evolution. Transportation Research Record 2254, 44-57.

Paleti, R., Pendyala, R.M., Bhat, C.R., Konduri, K.C., 2012. A Joint Tour-Based Model of Tour Complexity, Passenger Accompaniment, Vehicle Type Choice, and Tour Length. In: Proceedings of 91st Annual Meeting of the Transportation Research Board, National Research Council D.C.

Phaneuf, D.J., Kling, C.L., Herriges, J.A., 2000. Estimation and Welfare Calculations in a Generalized Corner Solution Model with an Application to Recreation Demand. Review of Economics and Statistics 82, 83-92.

Phaneuf, D.J., Smith, V.K., 2005. Recreation Demand Models. Handbook of Environmental Economics 2, K. North-Holland.

Pinjari, A.R., 2011. Generalized Extreme Value (GEV)-Based Error Structures for Multiple Discrete-Continuous Choice Models. Transportation Research Part B 45 (3), 474-489.

Pinjari, A.R., Bhat, C.R., 2010. A Multiple Discrete-Continuous Nested Extreme Value (MDCNEV) Model: Formulation and Application to Non-Worker Activity Time-Use and Timing Behavior on Weekdays. Transportation Research Part B 44 (4), 562-583.

Santos, A., McGuckin, N., Nakamoto, H.Y., Gray, D., Liss, S., 2011. Summary of Travel Trends: 2009 National Household Travel Survey. Publication FHWA-PL-ll-022. FHWA, U.S. Department of Transportation.

Schwarz, Gideon E., 1978. Estimating the Dimension of a Model. Annals of Statistics 6 (2), 461–464.

Srinivasan, S., Bhat, C.R., 2005. Modeling Household Interactions in Daily In-Home and Out-of-Home Maintenance Activity Participation. Transportation 32 (5), 523-544.

von Haefen, R.H., 2003. Incorporating Observed Choice into the Construction of Welfare Measures from Random Utility Models. Environmental Economics & Management 45 (2), 145-165.

von Haefen, R.H., Phaneuf, D.J., 2005. Continuous Demand System Approaches to Nonmarket Valuation. In: Applications of Simulation Methods in Environmental & Resource Economics, Springer, Dordrecht.

Wales, T.J., Woodland, A.D., 1983. Estimation of Consumer Demand Systems with Binding Non-negativity Constraints. Elsevier 21(3), 263-285.

LIST OF TABLES

TABLE 1: Model Fitness Measures

Models Model Fitness Criteria	MDCEV	Latent MDCEV With Two Segments	Latent MDCEV With Three Segments	Latent MDCEV With Four Segments
Number of Parameters	73	111	140	206
Log Likelihood	-17082.69	-16528.46	-16283.86	-16362.56
Bayesian information criterion (BIC)	34724.06	33906.40	33639.13	34301.63
Akaike information criterion(AIC)	34311.39	33278.92	32847.71	33137.11
Akaike information criterion Correction(AICc)	34316.70	33291.38	32867.79	33182.00

TABLE 2: Comparison of Latent Segmentation MDCEV Three Model with Exogenous Segmentation Based MDCEV Models

Segmentation Variables		N	Sub-dataset MDCEV Log Likelihood	Total MDCEV Log Likelihood
Gender	Male	840	-7456.2	-17033.4
	Female	1097	-9577.2	
Age	Under 21 years	207	-1670.9	-17047.9
	Over 21 years	1730	-15377.0	
Household size	<3 ppl	1068	-8684.6	-17035.6
	≥3 ppl	869	-8351.0	
Residential density	≤10K/sq mi	1517	-14047.7	-17012.6
	>10K/sq mi	420	-2965.0	
Season	Spring	394	-3564.4	-17047.4
	Not Spring	1543	-13483.0	
Gender and Age Classification	Males Under 21 years	109	-879.8	-16966.6
	Males Over 21 years	731	-6536.7	
	Female Under 21 years	98	-760.0	
	Female Over 21 years	999	-8790.2	
Gender and Household size	Male * Household Size <3	491	-4101.9	-16958.5
	Male * Household Size ≥3	349	-3316.1	
	Female * Household Size <3	577	-4547.0	
	Female * Household Size ≥3	520	-4993.5	
Three-segment MDCEV	A function of Male, Age under 21 years, Household size, Residential density >10K/sq mi, and Spring	1937	-	-16283.86

TABLE 3: Effects of Exogenous Variables on Segmentation Baseline Performance in the Latent Segmentation MDCEV Three

Explanatory Variables (Segment 1 is base)		Segment Two (Segment 1 is base)		Segment Three	
		Parameter	t-stat	Parameter	t-stat
Individual Characteristics	Male	-	-	0.34	1.96
	Age less than 21years	0.51	2.32	-	-
Household Characteristics	Household size	0.17	3.18	-	-
	Residential density>10K/sq mi	-	-	-0.83	-2.96
Contextual Variables	Spring	-0.86	-3.62	-	-
Constant		-1.36	-7.53	-0.97	-7.14
Number of cases		1937			
Log Likelihood at convergence		-16283.86			

TABLE 4: Latent Segmentation MDCEV Three Segment Characteristics

Segmentation Variables / **Share (%)**		**Segment One**	**Segment Two**	**Segment Three**	**Sample (Total)**
Male		41.51	41.81	49.82	43.37
Age less than 21years		8.78	17.17	9.05	10.69
Household size	<3 ppl	57.79	46.07	57.43	55.14
	≥3 ppl	42.21	53.93	42.57	44.86
Residential density >10K/sq mi		24.29	24.26	12.25	21.68
Spring		22.58	11.87	23.12	20.34

TABLE 5A: Effects of Exogenous Variables on Segment One Baseline Performance in the Latent Segmentation MDCEV Three

		Individual Socio-demographics					Household (HH) Socio-Demographics							
		Male	White	University Degree	22≤Age <60	Age≥60	HH size	Kids age ≤5 yrs	Kids age 6-15 yrs	Kids age 16-21 yrs	Income 40k - 70k	Income ≥ 70K	Urban area	Residential density >10K/sq mi
Activity Purpose Dimension (Baseline: Shopping)	Social/Recreational	-	0.52 (3.02)	-	-1.69 (-6.05)	-1.72 (-6.11)	-	-	-	-	-	0.33 (2.65)	-	-
	Transport Someone	-	-	0.74 (2.98)	-	-	0.30 (2.87)	3.06 (6.60)	2.97 (6.92)		-	-	-0.59 (-2.12)	-1.62 (-3.65)
	Meals	-	0.53 (1.34)	-	-1.24 (-3.40)	-1.48 (-4.08)	-	-	-	-	-	-	-	-
	Others	-	-	-	-1.21 (-3.75)	-1.33 (-4.11)	-	-	-	-	-	-	-	-
Accompaniment Dimension (Baseline: Alone)	With Household Member	-	-	-0.64 (-5.66)	-	-	-	1.45 (9.89)	1.20 (8.96)	-	-	-	-	-1.44 (-8.76)
	With HH & non-HH Member	-	-	-	-2.77 (-10.40)	-3.82 (-9.82)	-	-	-	-	-	-	-	-
Travel Mode Dimension (Baseline: Car)	Van/ Other Vehicles	1.40 (4.88)	-	-	-	-	0.96 (9.28)	-	2.22 (8.19)	-	-	-	-	-
	SUV	1.64 (8.97)	-	-	-	-	-	-	-	-	-	-	-	-
	Transit	-	-	-	-	-1.15 (-2.88)	0.46 (5.33)	-	-	-	-0.55 (-2.17)	-	-	4.11 (13.73)
	Walk/Bike	-	-	-	-1.04 (-2.85)	-	0.36 (6.45)	-	-	-	-	-	-	2.78 (14.22)

TABLE 5B: Effects of Exogenous Variables on Segment Two Baseline Performance in the Latent Segmentation MDCEV Three

		Individual Socio-Demographics			Household (HH) Socio-Demographics			Contextual Variables			
		Male	22<Age<60	Age≥60	HH size	Kids age ≤5 yrs	Kids age 6-15 yrs	Summer	Spring	Winter	Friday
Activity Purpose Dimension (Baseline: Shopping)	Social/Recreational	-	-	-	-	-	-	0.53 (2.25)	-	-	-
	Transport Someone	-	1.74 (4.48)	1.57 (3.64)	-0.45 (-4.49)	-	1.67 (5.98)	-	1.17 (3.93)	0.71 (3.65)	-0.93 (-3.15)
	Meals	-	-	-	-	-	-	-	-	-	-
	Others	-	-	-	-	-	-	-	-	-	-
Accompaniment Dimension (Baseline: Alone)	With Household Member	-	-1.84 (-6.70)	-	-	2.79 (6.29)	-	-	-3.24 (-9.34)	-	-
	With HH & non-HH Member	-	-1.40 (-5.22)	-	0.42 (6.21)	-	-	-	-	-	-
Travel Mode Dimension (Baseline: Car)	Van/ Other Vehicles	-	-	-	-	9.38 (4.71)	-	-	-	-	-
	SUV	-1.84 (-3.21)	-	-2.83 (-4.76)	-	-	-	-	-	6.67 (3.03)	-5.44 (-6.63)
	Transit	-	-	-	-	-	-	-	-	-	-
	Walk/Bike	-	-	-	-	-	-	-	-	-	-

TABLE 5C: Effects of Exogenous Variables on Segment Three Baseline Performance in the Latent Segmentation MDCEV Three

		Individual Socio-Demographics			**Household (HH) Socio-Demographics**						**Contextual Variables**	
		Male	University Degree	Age≥60	HH size	Kids age ≤5 yrs	Kids age 6-15 yrs	Income 40k -70k	Income ≥ 70K	Urban area	Spring	Friday
Activity Purpose Dimension (Baseline: Shopping)	Social/Recreational	-	-	-1.78 (-5.45)	-	-	-	-1.54 (-4.71)	-	-	-	-
	Transport Someone	-	-	-	0.35 (4.69)	-	-	-	-	-	-	-
	Meals	-	-	-	-	-	-	-	-	-	-	-
	Others	-	-	-	-	-	-	-	-	-	-	-
Accompaniment Dimension (Baseline: Alone)	With Household Member	-	0.62 (2.65)	1.55 (6.46)	-	1.15 (2.28)	-	-	-	-0.96 (-4.22)	1.98 (7.96)	-
	With HH & non-HH Member	-0.97 (-5.33)	-	-	-	-	-	-	-	-	-	-
Travel Mode Dimension (Baseline: Car)	Van/ Other Vehicles	-	-	-	-1.22 (-6.63)	-	-	-	-	-	1.86 (3.15)	-
	SUV	-3.18 (-4.64)	-	-	-	-	3.14 (5.50)	-	-	-	-	1.82 (3.97)
	Transit	-	-	-1.72 (-3.97)	-	-	-	-	-1.42 (-3.56)	-	-	-
	Walk/Bike	-	-	-3.12 (-6.73)	-	-	1.00 (3.26)	-2.38 (-5.87)	-	-	-	-

TABLE 6: Validation Results for the 2009 NHTS Sample

		Accompaniment Dimension			Activity Purpose Dimension					Travel Mode Dimension				
		Alone	With Household Member	With HH & non-HH Member	Shopping	Social/ Recreational	Transport Someone	Meals	Others	Car	Van/ Other	SUV	Transit	Walk/Bike
Observed	Participation	1365	744	370	1198	883	325	448	634	990	211	370	200	583
	Participation Rate (%)	70.47	38.41	19.10	61.85	45.59	16.78	23.13	32.73	51.11	10.89	19.10	10.33	30.10
MDCEV	Predicted Participation Rate (%)	67.11	49.70	24.48	48.95	50.13	18.73	20.03	30.08	57.47	13.96	25.38	12.33	28.50
	Percentage Error (%)	-4.77	29.40	28.18	-20.85	9.96	11.65	-13.40	-8.11	12.43	28.15	32.86	19.40	-5.31
	RMSE	**9.55**												
	MAE (%)	**17.27**												
Two-segment MDCEV	Predicted Participation Rate (%)	67.29	47.09	22.90	50.48	49.76	16.93	20.54	30.99	56.36	13.28	23.45	13.12	28.81
	Percentage Error (%)	-4.52	22.60	19.90	-18.37	9.16	0.91	-11.18	-5.33	10.27	21.87	22.76	27.02	-4.27
	RMSE	**7.85**												
	MAE (%)	**13.71**												
Three-segment MDCEV	Predicted Participation Rate (%)	67.23	45.94	21.77	51.59	49.83	17.51	20.80	30.99	54.79	11.87	22.63	13.25	28.38
	Percentage Error (%)	-4.60	19.60	13.95	-16.59	9.31	4.36	-10.07	-5.32	7.20	8.96	18.45	28.36	-5.72
	RMSE	**6.89**												
	MAE (%)	**11.73**												

TABLE 7: Validation Results of the Hold Out Sample for the 2009 NHTS Dataset

		Accompaniment Dimension			**Activity Purpose Dimension**					**Travel Mode Dimension**				
		Alone	With Household Member	With HH & non-HH Member	Shopping	Social/ Recreational	Transport Someone	Meals	Others	Car	Van/ Other	SUV	Transit	Walk/Bike
Observed	Participation	264	142	85	242	156	76	81	129	198	34	78	34	125
	Participation Rate (%)	69.84	37.57	22.49	64.02	41.27	20.11	21.43	34.13	52.38	8.99	20.63	8.99	33.07
MDCEV	Predicted Participation Rate (%)	67.08	49.51	24.04	48.79	49.18	19.18	19.55	29.89	57.38	10.87	25.58	12.51	29.71
	Percentage Error (%)	-3.96	31.79	6.92	-23.80	19.16	-4.59	-8.77	-12.41	9.54	20.85	23.97	39.03	-10.17
	RMSE	**2.01**												
	MAE (%)	**16.54**												
Two-segment MDCEV	Predicted Participation Rate (%)	67.21	47.07	22.27	50.20	48.94	17.37	20.23	30.68	55.91	10.79	23.27	13.36	30.16
	Percentage Error (%)	-3.77	25.31	-0.98	-21.58	18.57	-13.61	-5.58	-10.11	6.74	19.92	12.77	48.54	-8.80
	RMSE	**1.76**												
	MAE (%)	**15.10**												
Three-segment MDCEV	Predicted Participation Rate (%)	67.10	45.86	21.03	51.37	48.94	17.74	20.47	30.64	54.84	9.76	22.79	13.50	29.52
	Percentage Error (%)	-3.92	22.09	-6.47	-19.76	18.58	-11.76	-4.47	-10.21	4.70	8.47	10.45	50.05	-10.74
	RMSE	**1.63**												
	MAE (%)	**13.98**												