Transportmetrica A: Transport Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ttra21

A latent class model with fuzzy segmentation and weighted variables

Robert Ishaq[a], Shlomo Bekhor[a] & Yoram Shiftan[a]
[a] Transportation and Geo-Information Department, Faculty of Civil and Environmental Engineering, Technion - Israel Institute of Technology, Haifa, Israel
Published online: 10 Apr 2014.

Transportmetrica A: Transport Science, 2014
http://dx.doi.org/10.1080/23249935.2014.892543

Taylor & Francis
Taylor & Francis Group

A latent class model with fuzzy segmentation and weighted variables

Robert Ishaq*, Shlomo Bekhor and Yoram Shiftan

Transportation and Geo-Information Department, Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa, Israel

(*Received 8 May 2011; accepted 5 February 2014*)

Latent class models (LCMs) can yield powerful improvements in understanding the travel behaviour over traditional approaches. All the LCMs studies in transportation used discrete choice models for both the choice model and the identification of segment membership. This paper introduces an innovative segmentation methodology for the segment (class) identification model. The method includes a fuzzy segmentation process, which takes into account the varying levels of influence of each attribute on the degree of association with a segment. Five mode choice models were estimated using a data set from a household survey: a multinomial logit model, a nested logit (NL) model, a traditional LCM, a LCM using new segment identification, and a mixed NL. The estimation results indicate that the new segmentation method used for LCM captures heterogeneity differently than the traditional models, with similar likelihood estimates and good prediction results.

Keywords: latent class model; mixed logit model; fuzzy segmentation

1. Introduction

Latent class models (LCMs) are used to identify segments (classes) of respondents who tend to have similar preferences. LCMs classify respondents into different segments and estimate the utilities for each segment. In other words, the underlying theory of LCM posits that individual behaviour depends on observable attributes and on latent heterogeneity that varies with factors that are unobserved by the analyst.

Objects belonging to the same segment are similar with respect to the observed variables in the sense that their observed scores are assumed to come from the same probability distributions whose parameters are, however, unknown quantities to be estimated.

LCM assumes that a discrete number of segments (classes) are sufficient to account for preference heterogeneity across segments. Therefore, the unobserved heterogeneity is captured by these latent classes in the population, each of which is associated with a different parameter vector in the corresponding utility (Dillon and Kumar 1994; Wedel and Kamakura 2000).

The basic form of the LCM as presented by Bhat (1997) in an implementation for mode choice is composed of two components: (1) the probability that an individual i chooses mode j from the set C_i of available alternatives, conditional on the individual belonging to segment s, takes the familiar multinomial logit (MNL) form

$$P_i(j|s) = \frac{e^{\beta_s' x_{ij}}}{\sum_{w \in C_i} e^{\beta_s' x_{iw}}}, \tag{1}$$

*Corresponding author. Email: roberti@technion.ac.il

where x_{ij} is a vector of level-of-service and alternative-specific variables associated with alternative j and individual i and β_s is a parameter vector to be estimated. (2) The probability that individual i belongs to segment s is next written as a function of a vector z_i of socio-demographic and trip-related variables associated with the individual (z_i includes a constant). Using a MNL formulation

$$P_{is} = \frac{e^{\gamma_s' z_i}}{\sum_l e^{\gamma_l' z_i}}. \tag{2}$$

The unconditional (on segment membership) probability of individual i choosing mode j from the set C_i of available alternatives can be written from Equations (1) and (2) as follows:

$$P_{(ij)} = \sum_{s=1}^{S} P_{is} \times P_i(j|s). \tag{3}$$

Over the years researchers have realised that the use of LCM can yield powerful improvements in model explanation over traditional approaches. In transportation, there has been a wide use of LCM to better understand the travel behaviour. All the LCMs studies in transportation reviewed by the authors used discrete choice models for the choice model $P(j|s)$, mostly MNL (for example, Gopinath 1995; Walker and Li 2007) and MNL formulation to identify segment membership P_{is} (Swait 1994; Gopinath 1995; Bhat 1997; Boxall and Adamowicz 1999, 2002; Walker and Ben-Akiva 2002; Lee et al. 2003; Lee and Timmermans 2007; Walker and Li 2007; Beckman and Goulias 2008; Hess et al. 2009; Zhang et al. 2009; Hensher and Greene 2010; Hess, Shires, and Bonsall 2013; Atasoy, Glerum, and Bierlaire 2011). In the segment identification model P_{is} (see Equation (2)) each individual is probabilistically assigned to a segment based on individual-related characteristics using MNL mode. The sum of all probabilities is equal to one. The segment identification process is also known as fuzzy segmentation as individuals have partial membership in more than one segment (see Wedel and Kamakura 2000).

LCMs are widely used in many other areas such as marketing (Swait and Adamowicz 2001) and social sciences (Jürgen and Rolf 1997) and there are number of software tools which analyse and/or estimate LCM (for example, latent GOLD, Vermunt and Magidson 2000; BIOGEME, Bierlaire 2008; Mplus 2010; PROC LCA, Lanza et al. 2007). Other studies discussing segment identification models (grade-of-membership) incorporated into LCMs can be found in Varki, Cooil, and Rust (2000).

LCMs and mixed logit models (ML) (McFadden and Train 2000) try to better capture unobserved heterogeneity among individuals. In the ML model, the probability of choosing an alternative is calculated by integrating over all possible taste values. This integration requires assumptions about the structure and distribution of the tastes (e.g. normal, log-normal, etc.); once this structure is assumed, the probability can be estimated (Train 2003). In a LCM, the distributions are pre-specified (e.g. according to socio-economic or latent variables), thus enabling an estimation of the combined distribution.

A number of studies compare the two models (Greene and Hensher 2003; Shen, Sakata, and Hashimoto 2006; Hess et al. 2009). The main conclusions of these studies are that both models offer alternative ways of capturing unobserved heterogeneity and other potential forms of variability in unobserved sources of preferences. Furthermore, the final likelihood considering the number of estimated parameters indicates that the two models provide very close estimation results although each has its own merits. The LCM has the virtue of being a semi-parametric specification, which frees the analyst from possibly strong or unwarranted distributional assumptions about individual heterogeneity. The ML, while fully parametric, is sufficiently flexible to provide the modeller with a wide range in order to specify individual, unobserved heterogeneity. Both models usually improve explanatory power significantly compared to the MNL model and allow the analyst to harvest a rich variety of information about behaviour.

The objective of this paper is to introduce a new segmentation methodology for LCM used in transportation field in order to obtain an improved LCM explanatory power and estimation results. Towards this end, this methodology was compared to traditional LCM and other models.

A note on terminology is that: LCMs as defined in this paper are a form of ML models where the mixture distribution is assumed to be discrete, while ML models are more general than merely as a tool to capture random taste heterogeneity. For consistency with papers of related subjects, we keep the reference as LCM.

2. Methodology

The LCM in this study employs an innovative segmentation methodology for the segment (class) identification model (P_{is} – see Equation (2)) that includes a fuzzy segmentation process and takes into account the varying levels of influence of each attribute on the degree of association with a segment.

The segment identification model in this study is based on fuzzy methods, which assume that an object can belong to more than one segment. Its degree of association with a segment (P_{is} as in Equation (2)) is represented by a number in the interval [0,1] and the total association of an object with a segment must add up to one. A popular, well-known method of fuzzy segmenting is the 'Fuzzy C-means' (FCM) (Dunn 1974; Bezdek 1974), in which the degree of association with a segment is represented by the distance measured between the data point x_i and the segment centre c_s

$$u_{is} = \frac{1}{\sum_{h=1}^{S} ((x_i - c_s)/(x_i - c_h))^{2/(m-1)}}, \tag{4}$$

where u_{is} is the degree of association with x_i in segment s; i is Person I; x_i is a vector of attributes of person i; m is the fuzzifier, any real number greater than 1 (usually 2); $x_i - c_s$ is the distance $[d(c_s, x_i)]$ measured between the data point x_i and the segment centre c_s, see Equation (5); S is the number of segments $h = 1, \ldots, s, \ldots, S$ and c_h is the segment centre h.

Parameter m is called fuzzifier. For $m \to 1$, we have for the membership degree $u_{is} \to 0/1$, so the classification tends to be crisp. If $\to \infty$, then the $u_{is} \to 1/s$, where s is the number of clusters.

In the C-means segmentation process, all variables are treated as if they have the same weight in determining the degree of association with an object. In practical terms, for data where few variables determine particular segments, other variables may disguise the structure and therefore should not be considered in determining the degree of association for these segments. Taking into account the varying levels of influence of each attribute on the degree of association with a segment can be done by weighting the attributes for each segment as shown by Keller and Klawonn (2000) and presented in Equation (5), in which the new distance measure between a datum x_i and a segment (vector) c_s is defined by

$$d^2(c_s, x_i) = \sum_{k=1}^{K} \alpha_{sk}^{t} \cdot \left(x_i^{(k)} - c_s^{(k)}\right)^2, \tag{5}$$

where $x_i^{(k)}$ and $c_s^{(k)}$ indicate the kth coordinates of the vectors x_i and c_s, respectively.

The number of variables or attributes is denoted by k. α_{sk}^{t} is a parameter determining the weight of attribute k on the belonging to segment s, in other words it gives information of the influence of a variable or an attribute k of the data set on cluster s. S is the number of segments ($s = 1, \ldots, S$).

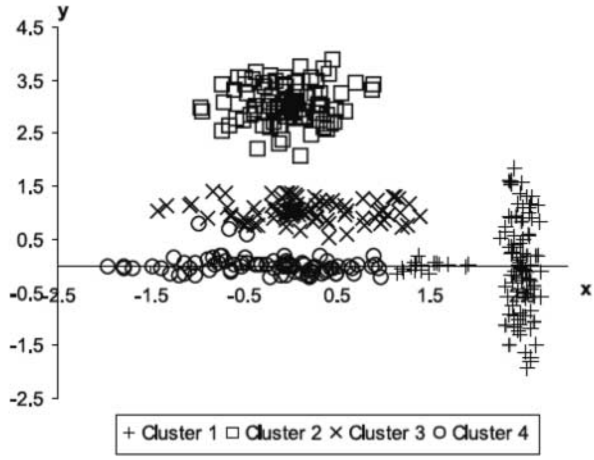

Figure 1. Results for ellipsoidal segments with FCM.
Source: Keller and Klawonn (2000).

Table 1. Attribute weights for ellipsoidal data set.

	Attributes	
Clusters	α_{sx}	α_{sy}
Cluster 1	0.99	0.01
Cluster 2	0.49	0.51
Cluster 3	0.08	0.92
Cluster 4	0.01	0.99

Source: Keller and Klawonn (2000).

Parameter α_{sk} can take any number between 0 and 1, higher value means higher influence. The parameter α_{sk} can be considered fixed or adapted individually for each segment during segmentation. The exponent t has a similar influence on the parameters α_{sk} as the fuzzifier m on the membership degree u_{is}. For $t \to 1$, the α tends to be 1 or 0 – either one attribute has unrestricted influence or no influence at all. On the other hand, if $t \to \infty$, all attributes get the same influence on the cluster structure, i.e. $\alpha_{sk} \to 1/\alpha$ for all k and s.

The segmentation method adopted from Keller and Klawonn (2000) has not been used in transportation neither to analyse behaviour, and is shown in the following example (adapted from Keller and Klawonn), where the data are segmented into four segments as shown in Figure 1. Table 1 gives the attribute weights α_{sk} for an ellipsoidal data set. It is obvious that for each segment, the more the data coordinates are scattered around those of the corresponding prototype, the less weight the corresponding attribute has for that segment. In this example, the two attributes influencing parameters α_{sk} for segment 2 have nearly the same value, as the data coordinates are approximately uniformly distributed for the two domains. For segments 3 and 4, the data values for attribute x are scattered widely, whereas the values for attribute y have a small range; thus, the influence of parameters α_{sx} is smaller than that of α_{sy} for segments 3 and 4. In the case of segment 1, the data values for attribute y are scattered widely, resulting in a high influence value for parameter α_{1x}.

Unlike LCMs presented in the transportation research, the LCM in this paper is based on the segmentation method of Keller and Klawonn (2000), which uses the C-means algorithm allowing the variables to have different weights in determining the degree to which an individual belongs to a particular segment. This is critical when the data are collected from a heterogeneous population in which the importance of each variable in determining the degree to which an individual belongs to a segment varies. The fragmented association of each individual belonging to a segment determines the characteristics of that segment. These segments are latent classes, since their characteristics are not directly observed but are inferred from other variables that are observed and directly measured.

A major issue in the segmentation process is the determination of the appropriate number of segments. A large number of methods have been proposed to identify the best number of segments. The most popular approach in practice is the 'Gap Statistics' method (Tibshiri et al. 2000). The idea behind the method is that the best number of segments is determined so that adding another segment does not improve much the modelling fit of the data. The method provides a plot of the number of segments (axis X) against a segment criterion (axis Y), with the best number of segments for a given data being determined where there is a 'bend' of the curve.

Another method for determining the number of segments is a statistical approach (Boxall and Adamowicz 2002) that penalises the improvement in the log-likelihood values as additional segments are added to the model. Thus, following Kamakura and Russell (1989), Gupta and Chintagupta (1994), and Swait (1994), two criteria are used to assist in determining the number of the segments: the minimum Akaike Information Criterion (AIC) and the minimum Bayesian Information Criterion (BIC).

In this research we estimate the LCM with the new methodology and compare it to the MNL and the nested logit (NL) as baseline models and to traditional LCM with segment identification using the MNL model (LCM–MNL) presented in transportation field and to mixed-NL as a mixture models.

3. Empirical implementation

Five models were estimated in the empirical implementation: (1) the MNL, (2) the NL, (3) the LCM with segment identification using the MNL model (LCM–MNL), (4) the LCM with segment identification using fuzzy segmentation c-mean algorithm with weighted variables, LCM-Weighted C-Mean (LCM–WCM), and (5) the mixed NL. The following sections describe the data source, the model definition, and the estimation results.

3.1. *Data source*

The Haifa travel habit survey, conducted by the Yefe Nof Company between 2005 and 2006 (Yefe Nof 2006), was used to estimate a mode choice model. Haifa is the largest city in northern Israel and the third largest metropolitan area in Israel.

In order to simplify the model representation and to focus on the model structure, it was decided to limit the number of alternatives and explanatory variables. Only trips that were carried out by the driver of a car, a passenger on a bus, or by walking were used, resulting in a sample of 12,754 trips. For each trip, five characteristics were selected for model estimation: gender, number of cars per household, total car travel time, total bus travel time, and walking travel time. The data analyses, given in Table 2, show that 49% of the trips were carried out by car, 26% by bus, and 25% by walking. The average time of these trips were 11.5 min by car, 31.5 min by bus, and 14.3 min by walking.

3.2. *Model specification*

Following the 'Gap Statistics' method (Tibshirani, Walther, and Hastie 2001), the data in the LCM models were divided into three segments. The models specifications are presented as follows:

Table 2. Descriptive statistics.

	Gender		Number of cars per household				
Alternative choice	Men	Women	0	1	2	3	Total
Driver	28%	21%	0%	30%	17%	2%	49%
Bus	10%	16%	16%	8%	2%	0%	26%
Walk	11%	14%	10%	12%	2%	0%	25%
Total	49%	51%	26%	50%	22%	2%	100%

Total travel time, by the alternative chosen (min)

	Car	Bus	Walk
Average travel	11.5	31.3	14.2
Minimum	1	4	5
Maximum	75	150	40

3.2.1. *MNL model*

The MNL probability function is presented in Equation (6)

$$P_{ij} = \frac{e^{Vij}}{\sum_J e^{Vij}}, \tag{6}$$

where P_{ij} is the probability of individual i to choose alternative j and V_{ij} is the utility function of alternative. The utility functions of the three alternatives in the MNL model include nine parameters ($\beta_1, \ldots, \beta_9$), as follows:

$$V_{\text{Driver}} = \beta_1 + \beta_2 \cdot \text{Gender} + \beta_3 \cdot \text{Ncars} + \beta_4 \cdot \text{CTime}, \tag{7}$$

$$V_{\text{Bus}} = \beta_5 + \beta_6 \cdot \text{Gender} + \beta_7 \cdot \text{Ncars} + \beta_8 \cdot \text{BTime}, \tag{8}$$

$$V_{\text{Walk}} = \beta_9 \cdot \text{WTime}, \tag{9}$$

where gender = 1 men, 0 women, Ncars = number of cars per household 0,1,2,+3, CTime = car total travel time (min), BTime = bus total travel time (min), and WTime = walk total travel time (min).

Note that the 'Driver' alternative is available only for individuals holding a driving license and that the 'Walk' alternative is available only for individuals who can make their trips by walking in less than 60 min.

3.2.2. *NL specification*

The NL probability function is presented in Equation (10)

$$P_{ij} = \frac{e^{V_j/\lambda_k}\left(\sum_{w\in B_k} e^{V_w/\lambda_l}\right)^{\lambda_k - 1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} e^{V_w/\lambda_l}\right)^{\lambda_l}}, \tag{10}$$

where P_{ij} is the probability of individual i to choose alternative j, V_{ij} is the utility function of alternative j, and parameter λ_k measures the degree of independence in the unobserved factors among the alternatives in the same nest k.

The utility functions of the NL model are the same as in the MNL model, where 'Nest A' includes only the 'Driver' alternative and 'Nest B' includes both the 'Bus' and the 'Walk' alternatives. Accordingly, a total of 10 parameters were estimated.

3.2.3. *LCM–MNL specification*

In the LCM–MNL, the choice model is a NL model as in the previous section. The utility functions of the segment identification model, which is a MNL (following Equation'(2)), are defined as follows:

$$V_{\text{Segment 1}} = \delta_{11} \cdot \text{Gender} + \delta_{12} \cdot \text{Ncars} + \delta_{13} \cdot \text{CTime} + \delta_{14} \cdot \text{BTime} + \delta_{15} \cdot \text{WTime}, \quad (11)$$

$$V_{\text{Segment 2}} = \delta_{21} \cdot \text{Gender} + \delta_{22} \cdot \text{Ncars} + \delta_{23} \cdot \text{CTime} + \delta_{24} \cdot \text{BTime} + \delta_{25} \cdot \text{WTime}, \quad (12)$$

$$V_{\text{Segment 3}} = \delta_{31} \cdot \text{Gender} + \delta_{32} \cdot \text{Ncars} + \delta_{33} \cdot \text{CTime} + \delta_{34} \cdot \text{BTime} + \delta_{35} \cdot \text{WTime}. \quad (13)$$

In total, 45 parameters are estimated in the LCM–MNL model:

- [9 utility parameters (β_i) + 1 nest parameter (λ_k)] × 3 segments = 30
- 15 utility parameters (δ_{ij}) identifying the probability to belong to a segment

The probability function for the LCM–MNL model is given in Equation (3).

3.2.4. *LCM–WCM specification*

The choice model is a NL model as in the previous section. The segment identification model takes the form of Equations (4) and (5), where u_{ij} presents the degree of association with x_i in segment s. In this model, the sum of the weights α_{is}^t was not constrained to be 1, so α_{is}^t can receive any number between 0 and 1. Note that for interpretation purposes, what matters is the relative weight of α_{is}^t.

In general, parameter t can be estimated as it adds another degree of freedom. In our model, which has already a high degree-of-freedom, we set the t parameter to be 1, as in Keller and Klawonn (2000).

In total, 60 parameters are estimated in the LCM model:

- [9 utility parameters (β_i) + 1 nest parameter] × 3 segments = 30
- 5 variables identifying the segment centres (c_s) × 3 segments = 15
- 5 weights for each variable (α_{is}^t) × 3 segments = 15

Assuming that the probability of choosing a mode is given by the NL model, the probability function of the LCM–WCM model is given by the following expression:

$$P_{(ij)} = \sum_{s=1}^{S} u_{is} \times P_{(j|s)} = \sum_{s=1}^{S} \left[\frac{1}{\sum_{s=1}^{S} ((x_i - c_s)/(x_i - c_h))^{2/(m-1)}} \times \frac{e^{V_j/\lambda_k} (\sum_{w \in B_k} e^{V_w/\lambda_l})^{\lambda_k - 1}}{\sum_{l=1}^{K} (\sum_{j \in B_l} e^{V_w/\lambda_l})^{\lambda_l}} \right]. \quad (14)$$

3.2.5. *Mixed-NL specification*

The mixed-NL model specification is presented by the following equation (McFadden and Train 2000):

$$P_C = \int L_C(i; x; \beta) \times G(d\beta; \sigma), \quad (15)$$

where $G(d\beta; \sigma)$ is the cumulative distribution function, $C = (1, \ldots, J)$ is the choice set, x is a vector of the observed attributes, β is the random parameters, $L_C(i; x; \beta)$ is a NL model for choice set C (as in the previous section), and σ is a vector of parameters of the mixing distribution G. The utility functions of the three alternatives in the mixed ML model, which is implemented by simulation for the same data with 500 draws, contain only 11 parameters ($\beta_1, \dots, \beta_9, \sigma_1, \sigma_2$), as follows:

$$V_{\text{Driver}} = \beta_1 + \beta_2 \cdot \text{Gender} + (\beta_3 + \sigma_1) \cdot \text{Ncars} + \beta_4 \cdot \text{CTime}, \tag{16}$$

$$V_{\text{Bus}} = \beta_5 + \beta_6 \cdot \text{Gender} + (\beta_7 + \sigma_2) \cdot \text{Ncars} + \beta_8 \cdot \text{BTime}, \tag{17}$$

$$V_{\text{Walk}} = \beta_9 \cdot \text{WTime}, \tag{18}$$

where $\sigma_{1,2}$ is the standard deviation of normal distribution with mean of $\beta_{3,7}$.

As in the NL model, 'Nest A' in the MNL includes only the 'Driver' alternative and 'Nest B' includes both the 'Bus' and 'Walk' alternatives. It is important to mention that many mixed NL definitions were tested (parameter value drawn from a distribution was tested also for 'Gender' and 'Time') and that the definition above received the best estimation results.

3.3. *Estimation results*

As all models presented in this study are discrete choice models, the general log-likelihood function is

$$LL(\beta) = \sum_{i=1}^{I} \sum_{j} y_{ij} \ln P_{ij}, \tag{19}$$

where β is a vector of the parameters of the model, P_{ij} is the probability of individual i choosing alternative j, and y_{ij} is 1 if individual i chose j and 0 otherwise.

For example, the log-likelihood function of the LCM–WCM will be then as follows:

$$LL(\beta) = \sum_{i=1}^{I} \sum_{i} y_{ij} \ln \left[\sum_{s=1}^{S} P_{is} \times P_{(j|s)} \right] = \sum_{i=1}^{I} \sum_{i} y_{ij} \ln \left[\sum_{s=1}^{S} u_{is} \times P_{(j|s)} \right]$$
$$= \sum_{i=1}^{I} \sum_{i} y_{ij} \ln \left[\sum_{s=1}^{S} \left[\frac{1}{\sum_{s=1}^{S} ((x_i - c_s)/(x_i - c_h))^{2/(m-1)}} \times \frac{e^{V_{js}/\lambda_{ks}} (\sum_{w \in B_{ks}} e^{V_{ws}/\lambda_l})^{\lambda_{ks}-1}}{\sum_{l=1}^{K} (\sum_{j \in B_l} e^{V_{ws}/\lambda_{ls}})^{\lambda_{ls}}} \right] \right]. \tag{20}$$

The estimation process goal by simulation (Train 2003) is to find the value of β that maximises this function.

3.3.1. *MNL model*

The results of the estimated utility parameters of the MNL model are given in Table 3. First, all parameters are statistically significant. Second, all parameter signs are logical, showing that men are more likely to drive than women. Individuals from households with more cars are more likely to drive (see β_2, β_3) than use a bus (see β_5, β_6) and are even less likely to walk. Finally, all travel-time parameters are significant and negative (see $\beta_4, \beta_8, \beta_9$) and the walk travel-time coefficient has the highest absolute value, indicating that individuals are most sensitive to walk time. The final likelihood value is −7797.2.

3.3.2. *NL model*

The results of the estimated utility parameters of the NL model are given in Table 4. It can be seen that all the coefficients, including the nest coefficients, (λ_B), are statistically significant. As in the MNL model, all the parameters' signs are logical, with men and more cars per household

Table 3. MNL estimated utility parameters.

	Driver	Bus	Walk
Constant	−4.44 (−48.6)	−2.27 (−27.8)	–
Gender	0.76 (12.3)	−0.22 (−3.4)	–
Number of cars	1.95 (37.5)	−0.52 (−9.4)	–
Travel time	−0.038 (−6.3)	−0.014 (−5.7)	−0.107 (−39.5)

Value (t-test)
Observations: 12,754
Initial likelihood: −11,970.2
Final likelihood: −7797.2

Number of parameters: 9
Adjusted R^2: 0.348

Note: All coefficient values are statistically significant at the 5% significance.

Table 4. NL estimated utility parameters.

	Driver	Bus	Walk
Constant	−3.87 (−47.3)	−1.28 (−15.9)	–
Gender	0.80 (15.8)	−0.10 (−2.6)	–
Number of cars	1.93 (43.9)	−0.29 (−8.9)	–
Travel time	−0.030 (−5.2)	−0.023 (−10.6)	−0.068 (−21.7)

Nest A = 1.0, Nest B = 0.426(17.5)
Value (t-test)
Observations: 12,754
Initial likelihood: −11,970.2
Final likelihood: −7606. 1

Number of parameters: 10
Adjusted R^2: 0.364

Note: All coefficient values are statistically significant at the 5% significance.

(see β_2, β_3) increasing the utility of the 'Driver' alternative, and all travel-time parameters are significant and negative (see $\beta_4, \beta_8, \beta_9$).

According to the final likelihood value (−7606.1 for the NL compared to −7797.2 for the MNL), the likelihood ratio test (presented below), and the adjusted R^2 indicator (0.364 for NL compared to 0.348 for MNL), the NL model is significantly better than the MNL model.

$$\text{Likelihood ratio test}: \; -2 \times [(-7797.2) - (7606.1)] \cong 382 > \chi^2_{1.5\%} = 7.88.$$

3.3.3. *LCM–MNL*

The LCM–MNL estimates 45 parameters: the segment identification utility parameters (Table 5) and the choice utility parameters (Table 6). It is important to mention that in a segment identification model (which is MNL model), as opposed to mode choice model, time coefficients can receive a positive value. A positive time coefficient in the utility function indicates that the likelihood of choosing this segment increases with an increasing travel time.

Table 5. LCM–MNL segmentation estimated utility parameters.

	Gender	Number of cars	Walk travel time	Bus travel time	Walk travel time
Segment 1	5.75 (46.2)	0.41 (2.8)	0.001 (0.1)	0.334 (17.7)	0.235 (5.1)
Segment 2	6.22 (54.1)	0.75 (5.4)	0.000 (0.01)	0.363 (19.6)	0.229 (5.0)
Segment 3	6.56 (55.9)	0.41 (3.4)	−0.084 (−4.8)	0.382 (18.6)	0.314 (7.3)

Note: Underlined value (t-test) is insignificant at the 5% significance level.

Table 6. LCM–MNL choice estimated utility parameters.

		Driver	Bus	Walk
Segment 1	Constant	−6.87 (−22.2)	−3.59 (−5.0)	–
	Gender	−11.62 (−4.6)	−0.08 (−1.0)	–
	Number of cars	1.82 (7.9)	−0.04 (−0.8)	–
	Travel time	−0.132 (−3.7)	−0.004 (−0.7)	−0.100 (−4.9)
	Nest A = 1.0, Nest B = 0.162 (4.2)			
Segment 2	Constant	−12.93 (−9.1)	−0.25 (−3.6)	–
	Gender	8.67 (6.4)	−0.29 (−4.2)	–
	Number of cars	13.0 (9.3)	−0.42 (−5.9)	–
	Travel time	−0.101 (−7.2)	−0.005 (−7.6)	−0.064 (−8.1)
	Nest A = 1.0, Nest B = 0.291 (6.9)			
Segment 3	Constant	−12.46 (−7.4)	−1.26 (−5.9)	–
	Gender	9.93 (4.7)	−0.03 (−0.4)	–
	Number of cars	−1.17 (−1.8)	0.22 (4.8)	–
	Travel time	−0.280 (−2.3)	−0.010 (−1.0)	−0.057 (−4.4)
	Nest A = 1.0, Nest B = 0.171 (4.5)			

Value (*t*-test)
Observations: 12,754
Initial likelihood: −11,970.2
Final likelihood: −6877.6

Number of parameters: 45
Adjusted R^2: 0.422

Note: Underlined value (*t*-test) is insignificant at the 5% significance level.

Table 7. LCM–MNL segments' characteristics.

Variable	Segment 1	Segment 2	Segment 3
Gender	38% men, 62% women	51% men, 49% women	56% men, 44% women
Number of cars per HH	0.83	1.08	0.85
Travel time – car (min)	8.6	12.0	4.9
Travel time – bus (min)	23.4	33.3	12.5
Travel time – walk (min)	52.1	79.7	15.4
Actual mode choice*	D – 40%, B – 31%, W – 29%	D – 55%, B – 27%, W – 18%	D – 34%, B – 11%, W – 54%
Average degree of association	20%	68%	12%

*D-Driver, B-Bus, and W-Walk (average choice: 49%, 26%, and 25%, respectively).

According to Table 7, segment 1 includes 20% of the trips and is characterised as 38% men, 0.83 cars in household in average, and medium travel time (52 min by walk, 23 min by bus, and 9 min by car). Segment 2 includes 68% of the trips and is characterised as 51% men, 1.08 cars in household in average, and high travel time (79 min by walk, 33 min by bus, and 12 min by car). Segment 3 includes only 12% of the trips and is characterised as 56% men, 0.84 cars in household in average, and low travel time (15 min by walk, 12 min by bus, and 5 min by car). The LCM–MNL received a final likelihood value of −6877.6 (−2423.6 for segment 1, −2618.8 for segment 2, and −1835.2 for segment 3), a significant improvement compared to MNL and NL models.

According to the final likelihood value (−6877.6 for the LCM–MNL compared to −7606.1 for the NL and −7797.2 for the MNL), the likelihood ratio test (presented below), and the adjusted R^2 indicator (0.422 for the LCM–MNL compared to 0.364 for the NL and 0.348 for the MNL), the LCM–MNL model is significantly better than the NL and MNL models, despite the great addition in the estimated parameters.

$$\text{Likelihood ratio test}: -2\,[(-7797.2) - (-6877.6)] \cong 1839.2 > \chi^2_{35.5\%} = 48.6,$$

$$\text{Likelihood ratio test}: -2\,[(-7606.1) - (-6877.6)] \cong 1457.0 > \chi^2_{34.5\%} = 49.8.$$

3.3.4. *LCM–WCM*

The LCM–WCM estimates 60 parameters: segment centroid characteristics (Table 8), segment variable weights (Table 9), and utility parameters of the choice model (Table 10). According to Table 8, the segments are characterised according to 'gender', 'number of cars', 'travel time' variables, and their actual mode choice. The segments can be tagged according to the travel time, where segment 1 is labelled as 'long travel time', segment 2 'short travel time', and segment 3 'medium travel time'.

Overall, 53% of the trips belong to segment 1, 9% to segment 2, and 38% to segment 3. The weights of the variables are given in Table 9. The most influential variables on the degree of association with segment 1 are the 'number of cars per household' and 'bus travel time' (0.37 and 0.32, respectively); with segment 2, 'bus travel time' and 'walk travel time' (1.00 and 0.81, respectively); and with segment 3, 'number of cars per household', 'car travel time', and 'bus travel time' (0.31, 0.73, and 0.75, respectively).

The results given in Table 10 indicate that the 'Driver' alternative in segment 1 is less sensitive to car travel time (−0.07) compared with segments 2 and 3 (−0.152 and −0.125, respectively). On the other hand, the 'Bus' alternative in segment 1 is more sensitive to travel time than in segment 3 (−0.045 compared with −0.026), while segment 2 is the least sensitive to bus travel time (barely significant). All segments are very sensitive to walk travel time, which receive highly negative values (−0.188, −0.144, and −0.138, respectively).

Table 8. LCM–WCM segments — Centroids' characteristics.

Variable	Segment 1	Segment 2	Segment 3
Gender	58% men, 42% women	48% men, 52% women	37% men, 63% women
Number of cars per HH	1.1	1.8	0.72
Travel time – car (min)	10.8	7.1	8.4
Travel time – bus (min)	32.4	23.1	25.1
Travel time – walk (min)	77.4	45.5	56.1
Actual mode choice*	D – 58%, B – 19%, W – 23%	D – 63%, B – 13%, W – 24%	D – 34%, B – 40%, W – 26%
Average degree of association	53%	9%	38%

*D, Driver; B, Bus; W, Walk (average choice: 49%, 26%, and 25%, respectively).

Table 9. LCM–WCM segments — variable weights.

Variable	Segment 1	Segment 2	Segment 3
Gender	0.08	0.11	0.01
Number of cars per HH	0.37	0.63	0.31
Travel time – car (min)	0.01	0.55	0.73
Travel time – bus (min)	0.32	1.00	0.75
Travel time – walk (min)	0.01	0.81	0.01

Table 10. LCM–WCM choice estimated utility parameters.

		Driver	Bus	Walk
Segment 1	Constant	−5.12 (−30.5)	−0.77 (−8.9)	–
	Gender	0.81 (10.2)	−0.33 (−5.6)	–
	Number of cars	4.08 (29.5)	−0.60 (−9.8)	–
	Travel time	−0.070 (−6.7)	−0.045 (−11.2)	−0.118 (−14.3)
	Nest A = 1.0, Nest B = 0.261 (12.9)			
Segment 2	Constant	−5.86 (−15.6)	−2.72 (−9.6)	–
	Gender	−0.14 (−0.7)	0.001 (0.07)	–
	Number of cars	1.33 (9.2)	−0.55 (−4.1)	–
	Travel time	−0.152 (−4.3)	−0.018 (−1.7)	−0.144 (−10.7)
	Nest A = 1.0, Nest B = 1.0 (8.8)			
Segment 3	Constant	−9.93 (−23.0)	−3.8 (−13.9)	–
	Gender	−0.18 (−1.4)	−0.47 (−4.9)	–
	Number of cars	5.57 (20.4)	−0.89 (−9.6)	–
	Travel time	−0.125 (−6.0)	−0.026 (−4.2)	−0.138 (−15.6)
	Nest A = 1.0, Nest B = 0.366 (10.1)			

Value (*t*-test)
Observations: 12,754
Initial likelihood: −11,970.2
Final likelihood: −6855.0
Number of parameters: 60
Adjusted R^2: 0.422

Note: Underlined value (*t*-test) is insignificant at the 5% significance level.

According to the final likelihood value (−6855.0 for the LCM–WCM compared to −7606.1 for the NL and −7797.2 for the MNL), the likelihood ratio test (presented below), and the adjusted R^2 indicator (0.422 for the LCM–WCM compared to 0.364 for the NL and 0.348 for the MNL), the LCM–WCM model is significantly better than the NL and MNL models, despite the great addition in the estimated parameters.

$$\text{Likelihood ratio test}: -2\left[(-7797.2)-(-6855.0)\right] \cong 1884.4 > \chi^2_{50.5\%} = 67.5,$$
$$\text{Likelihood ratio test}: -2\left[(-7606.1)-(-6855.0)\right] \cong 1502.2 > \chi^2_{49.5\%} = 66.3.$$

3.3.5. *Mixed NL*

The estimated parameters of the mixed NL are given in Table 11. Three of the 12 parameters are insignificant at the 5% level of confidence. According to the final likelihood value (−7060.4 for the mixed-NL compared to −7606.1 for the NL and −7797.2 for the MNL), the likelihood ratio test (presented below), and the adjusted R^2 indicator (0.409 for the mixed-NL compared to 0.364 for the NL and 0.348 for the MNL), the mixed-NL model is significantly better than the NL and MNL models, despite the great addition in the estimated parameters.

$$\text{Likelihood ratio test}: -2\left[(-7797.2)-(-7060.4)\right] \cong 1473.6 > \chi^2_{2.5\%} = 5.99,$$
$$\text{Likelihood ratio test}: -2\left[(-7606.1)-(-7060.4)\right] \cong 1091.4 > \chi^2_{3.5\%} = 7.82.$$

The mixed NL received a final likelihood value of −7060.4, which is very close, but lower than the value of the LCM–MNL, −6877.6, and LCM–WCM, −6855.0. In this case, a likelihood ratio test cannot be applied as one model is not a special case of the other.

Table 11. Mixed NL estimated utility parameters.

	Driver	Bus	Walk
Constant	−8.62 (−13.1)	−2.22 (−9.9)	–
Gender	2.04 (4.1)	−0.42 (−1.44)	–
Number of cars	5.52 (24.2)	−1.62 (−1.0)	–
Travel time	−0.071 (−2.7)	−0.059 (−7.7)	−0.137 (−18.7)
Nest A = 1.0, Nest B = 0.82 (9.5)			
$\sigma_1 = \pm 3.2$ (8.3) $\sigma_2 = \pm 0.82$ (0.65)			
Value (*t*-test)		Number of parameters: 12	
Observations: 12,754		Adjusted R^2: 0.409	
Initial likelihood: −11,970.2			
Final likelihood: −7060.4			

Note: Underlined value (*t*-test) is insignificant at the 5% significance level.

Table 12. Comparison of models.

Model	No. of estimated parameters	Final likelihood	Adjusted R^2	AIC	BIC
MNL	9	−7797.2	0.348	15,612.4	15,679.5
NL	10	−7606.1	0.364	15,232.2	15,306.7
LCM–MNL	45	−6877.6	0.422	13,845.2	14,180.6
LCM–WCM	60	−6855.0	0.422	13,830.0	14,277.2
Mixed NL	12	−7060.4	0.409	14,144.8	14,234.2

3.3.6. *Model compression*

Table 12 compares the five models according to the different measures of model fit. In general, the models LCM–MNL, LCM–WCM, and mixed-NL improve significantly the estimation results compared to the baseline models, MNL and NL. In this study, the LCM models achieve very close but still higher likelihood value than mixed-NL, despite the increase in the number of the estimated parameters. Both LCM–MNL and LCM–WCM receive very close results regarding the likelihood value and the AIC measure. The BIC measure informs that the LCM–MNL has an advantage over the LCM–WCM and over the mixed-NL despite the small number of the estimated parameters (only 12 compared to 45 parameters in the LCM–MNL model).

In order to test the forecasting power of such models, we divided the data into two data set comprising 80% and 20% of the data randomly. The 80% (which contains 10,204 observations) were used to estimate the model parameters. The models with these new parameters were used to estimate the choice probability of the other 20% (2551 observations). The results, as given in Table 13, indicate that the final likelihood of the LCM–MNL, LCM–WCM, and Mixed-NL models for the 20% sample (−1524.5, −1518.3, and −1567.5, respectively) are significantly better compared with MNL and NL models (−1979.7, and −1925.6, respectively). The final likelihood of the LCM–MNL and LCM–WCM models for the 20% sample is very close and higher than the final likelihood of the MMNL model for the 20% sample, as was indicated in the estimation results of the full data. The prediction test ‘P-agree’ of the 20% sample, which indicates the match between the high-probability alternative and the actual choice, indicates that LCM–WCM achieved the best prediction result with 75.5% (1926 out of 2551), compared to LCM–MNL that achieved 74.9% (1911 out of 2551), and MMNL that achieved 74.7% (1906 out of 2551).

Table 13. Prediction power – final likelihood of 20% of the observations.

Model	Final likelihood	P-agree (out of 2551)
MNL	−1979.7	74.1% (1890)
NL	−1925.6	74.5% (1900)
LCM–MNL	−1524.5	74.9% (1911)
LCM–WCM	−1518.3	75.5% (1926)
Mixed NL	−1567.5	74.7% (1906)

4. Conclusions and future research

This study developed an innovative segment identification model methodology for LCM that includes a fuzzy segmentation process taking into account the different levels of influence that each attribute has on the degree of association with a segment.

The main contribution of the paper lies in the development and implementation of a segment identification model that considers the best variable mix for a given set of segmentation variables. As in LCM–MNL, the model does not treat all variables identically and the segmentation procedure presented here shows that not all variables have the same weight in determining segment association.

Both LCM–MNL and LCM–WCM were compared with a mixed NL model. As expected, all three models introduce significant gains in performance compared to the MNL and NL models. However, the estimation results in this case indicate that LCM–MNL and LCM–WCM capture the heterogeneity among individuals better than does the mixed NL. As reported in other studies (Greene and Hensher 2003; Shen, Sakata, and Hashimoto 2006; Hess et al. 2009), LCM is able to retrieve richer patterns of heterogeneity by linking segment allocation to socio-demographic and trip indicators. On the other hand, the simplicity of the mixed NL due to its small number of estimated parameters (12 parameters) compared to the large number of those estimated in the LCMs (45 for LCM–MNL and 60 for LCM–WCM) gives it an advantage especially when both models achieve close estimation results. That is to support what is brought in the findings of the studies of Greene and Hensher (2003), Shen, Sakata, and Hashimoto (2006), and Hess et al. (2009) that both models provide very close estimation results although each has its own merits and both models usually improve explanatory power significantly compared to the MNL model and allow the analyst to harvest a rich variety of information about behaviour.

Both LCMs presented in this work, the LCM–MNL and the LCM–WCM models, relay on different models for segment identifications. This fact yielded to different segment characteristics. Despite the differences in the segment compound, they both achieved close estimation results, which mean that they capture the unobserved heterogeneity of the population in a different way. This is to say that there are different heterogeneities in the same data set and they can be captured by applying different segmentation methods. Nonetheless, the prediction test indicates that the LCM–WCM model has an advantage on the LCM–MNL model.

The results presented in this paper illustrate the potential benefits of LCM for applied research in the area of travel behaviour, which guides the design and estimation of such a model towards more behaviourally realistic representations with an improved explanatory power.

This research presents that there are more than one segmentation methodology that capture the heterogeneity among a given data, in a different way, and achieve similar likelihood results. One of the directions that this research must take in the future is to investigate if there are other segment identifications, which capture the heterogeneity in a different way, that can achieve the same estimation results or better.

Acknowledgements

The authors wish to thank "Yefe Nof" company for providing the data for this research.

References

Atasoy, B., A. Glerum, and M. Bierlaire. 2011. "Mode Choice with Attitudinal Latent Class: A Swiss Case-Study." Proceedings of the second international choice modeling conference (ICMC), Leeds, July 4–6, 2011.

Beckman, J. D., and K. G. Goulias. 2008. "Immigration, Residential Location, Car Ownership, and Commuting Behavior: Multivariate Latent Class Analysis from California." Paper presented at the 87th transportation research board annual meeting, Washington, DC, January 13–17, and included in the CD ROM proceedings.

Bezdek, J. C. 1974. "Numerical Taxonomy with Fuzzy Sets." *Journal of Mathemetical Biology* 1 (1): 57–71.

Bhat, C. R. 1997. "An Endogenous Segmentation Mode Choice Model with an Application to Intercity Travel." *Transportation Science* 31 (1): 34–48.

Bierlaire, M. 2008. *An Introduction to BIOGEME Version 1.6.* biogeme.epfl.ch. Lausanne, Switerland: EPFL, École Polytechnique De Lausanne.

Boxall, P. C., and W. L. Adamowicz. 1999. "Understanding Heterogeneous Preferences in Random Utility Models: The Use of Latent Class Analysis." Staff Paper 99-02, Department of Rural Economy, University of Alberta Edmonton, Canada.

Boxall, P. C., and W. L. Adamowicz. 2002. "Understanding Heterogeneous Preferences in Random Utility Models: A Latent Class Approach." *Environmental and Resource Economics* 23 (4): 421–446.

Dillon, W. R., and A. Kumar. 1994. "Latent Structure and Other Mixture Models in Marketing: An Integrative Survey and Overview." In *Advanced Methods of Marketing Research*, edited by R. P. Bagozzi, 352–388. Cambridge: Blackwell Publishers.

Dunn, J. C. 1974. "A Fuzzy Relative of the ISODATA Process and Its Use in Determining Compact Well-Separated Clusters." *Journal of Cybernetics* 3 (3): 32–57.

Gopinath, D. 1995. "Modeling Heterogeneity in Discrete Choice Processes: Application to Travel Demand." PhD diss., Department of Civil Engineering, MIT, Cambrige, MA.

Greene, H., and A. Hensher. 2003. "A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit." *Transportation Research Part B: Methodological* 37 (8): 681–698.

Gupta, S., and P. K. Chintagunta. 1994. "On Using Demographic Variables to Determine Segment Membership in Logit Mixture Models." *Journal of Marketing Research* 31: 128–136.

Hensher, David, and William Greene. 2010. "Non-Attendance and Dual Processing of Common-Metric Attributes in Choice Analysis: A Latent Class Specification." *Empirical Economics* 39 (2): 413–426.

Hess, S., M. Ben-Akiva, D. Gopinath, and J. Walker. 2009. "Taste Heterogeneity, Correlation, and Elasticities in Latent Class Choice Models." 88th annual TRB meeting, Washington, DC.

Hess, S., J. D. Shires, and P. W. Bonsall. 2013. "A Latent Class Approach to Dealing with Respondent Uncertainty in a Stated Choice Survey for Fare Simplification on Bus Journeys." *Transportmetrica A: Transport Science* 9 (6): 473–493.

Jürgen, R., and L. Rolf. 1997. *Applications of Latent Trait and Latent Class Models in the Social Sciences*. New York: Waxmann Publishing Co.

Kamakura, W., and G. Russell. 1989. "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure." *Journal of Marketing Research* 26: 379–390.

Keller, A., and F. Klawonn. 2000. "Fuzzy Clustering with Weighting of Data Variables." *International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems* 8 (6): 735–746.

Lanza, S. T., L. M. Collins, D. R. Lemmon, and J. L. Schafer. 2007. "PROC LCA: A SAS Procedure for Latent Class Analysis." *Structural Equation Modeling* 14 (4): 671–694.

Lee, B., A. Fujiwara, J. Zhang, and Y. Sugie. 2003. "Analysis of Mode Choice Behaviours based on Latent Class Models." 10th international conference on travel behaviour research (IATBR). Lucerne, August 10–15.

Lee, B., and H. Timmermans. 2007. "A Latent Class Accelerated Hazard Model of Activity Episode Durations." *Transportation Research Part B: Methodological* 41 (4): 426–447.

McFadden, D., and K. Train. 2000. "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics* 15 (5): 447–470.

Mplus Software. 2010. http://www.statmodel.com
Shen, J., Y. Sakata, and Y. Hashimoto. 2006. "A Comparison Between Latent Class Model and Mixed Logit Model for Transport Mode Choice: Evidences from Datasets of Japan." Discussion Papers in Economics and Business 05–06, Osaka School of International Public Policy (OSIPP), Osaka University, Toyonaka, Osaka, Japan.
Swait, J. R. 1994. "A Structural Equation Model of Latent Segmentation and Product Choice for Cross-Sectional Revealed Preference Choice Data." *Journal of Retailing and Consumer Services* 1 (2): 77–89.
Swait, J., and W. Adamowicz, 2001. "The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching." *Journal of Consumer Research* 28: 135–148.
Tibshirani, R., G. Walther, and T. Hastie. 2001. "Estimating the Number of Clusters in a Data Set Via the Gap Statistic." *Journal of the Royal Statistical Society* 63 (2): 411–423.
Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambrige: Cambrige University Press.
Varki, Sajeev, Bruce Cooil, and Roland T. Rust. 2000. "Modeling Fuzzy Data in Qualitative Marketing Research." *Journal of Marketing Research* 37 (November): 480–489.
Vermunt, J. K., and J. Magidson. 2000. *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.
Walker, J., and M. Ben-Akiva. 2002. "Generalized Random Utility Model." *Mathematical Social Sciences* 43 (3): 303–343.
Walker, J. L., and J. Li. 2007. "Latent Lifestyle Preferences and Household Location Decisions." *Journal of Geographical Systems* 9 (1): 77–101. 12th WCTR, July 11–15, 2010, Lisbon, Portugal.
Wedel, M., and W. Kamakura. 2000. *Market Segmentation: Conceptual Methodological Foundations.* 2nd ed. Boston, MA: Kluwer Academic Publishers.
Yefe Nof. 2006. "Metropolitan Haifa Travel Habits Survey." Report. Haifa, Israel. http://www.yefenof.co.il/index.php?m_id=32&if_link=0
Zhang, J., M. Kuwano, B. Lee, and A. Fujiwara. 2009. "Modeling Household Discrete Choice Behavior Incorporating Heterogeneous Group Decision-Making Mechanisms." *Transportation Research Part B: A Special Issue 'Household Behavior Modeling'* 43: 230–250.