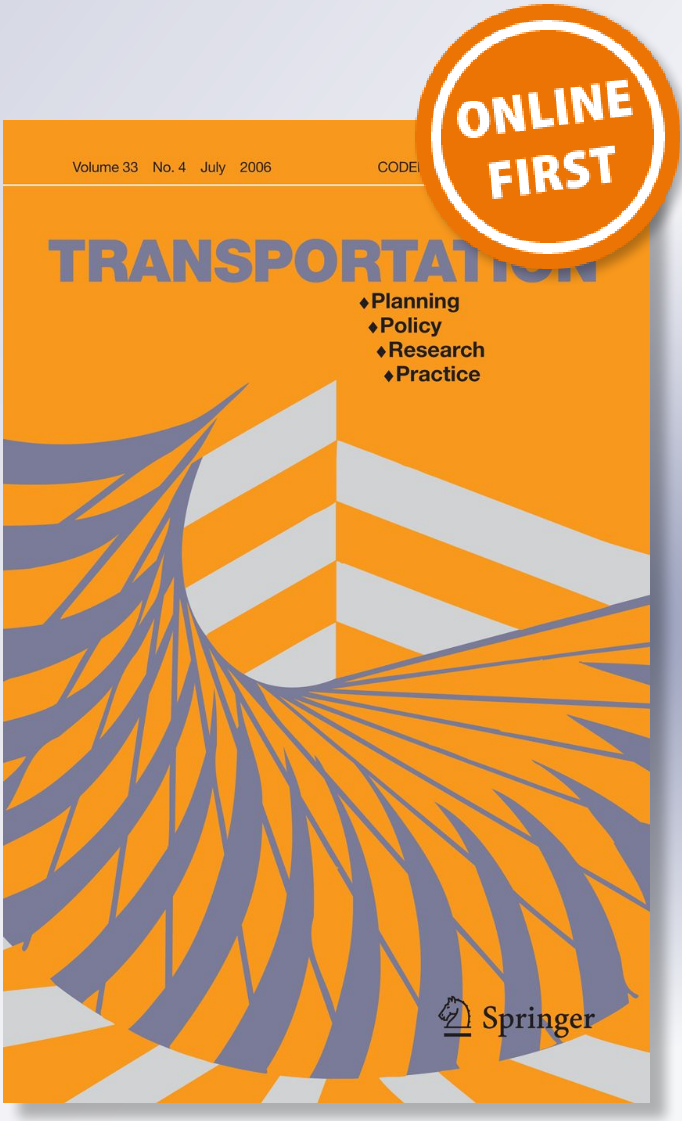
A flexible model structure approach for discrete choice models

Robert Ishaq, Shlomo Bekhor & Yoram Shiftan

Transportation
Planning - Policy - Research - Practice

ISSN 0049-4488

Transportation
DOI 10.1007/s11116-012-9431-8



Springer

Transportation
DOI 10.1007/s11116-012-9431-8

A flexible model structure approach for discrete choice models

Robert Ishaq · Shlomo Bekhor · Yoram Shiftan

Abstract Multi-dimensional discrete choice problems are usually estimated by assuming a single-choice hierarchical order for the entire study population or for pre-defined segments representing the behavior of an "average" person and by indicating either limited differences or a variety in choices among the study population. This study develops an integral methodological framework, termed the flexible model structure (FMS), which enhances the application of the discrete choice model by developing an optimization algorithm that segment given data and searches for the best model structure for each segment simultaneously. The approach is demonstrated here through three models that conceptualize the multi-dimensional discrete choice problem. The first two are Nested Logit models with a two-choice dimension of destination and mode; they represent the estimation of a fixed-structure model using pre-segmented data as is mostly common in multi-dimensional discrete choice model implementation. The third model, the FMS, includes a fuzzy segmentation method with weighted variables, as well as a combination of more than one model structure estimated simultaneously. The FMS model significantly improves estimation results, using fewer variables than do segmented NL models, thus supporting the hypothesis that different model structures may best describe the behavior of different groups of people in multi-dimensional choice models. The implementation of FMS involves presenting the travel behavior of an individual as a mix of travel behaviors represented by a number of segments. The choice model for each segment comprises a combination of different choice model structures. The FMS model thus breaks the consensus that an individual belongs to only one segment and that a segment can take only one structure.

Keywords Discrete choice models · Multi-Dimensional choice models · Segmentation · Logit

R. Ishaq (✉) · S. Bekhor · Y. Shiftan
Department of Transportation and Geo-Information, Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Technion Campus, 32000 Haifa, Israel
e-mail: roberti@technion.ac.il

Published online: 12 July 2012

Introduction

The attempt to understand and model travel behavior in a variety of situations has occupied many travel-demand researchers for years. Travel behavior lies at the core of procedures for analyzing and evaluating transportation-related measures that are aimed at improving urban and suburban mobility, environmental quality and a wide variety of social objectives.

Travel behavior has been estimated and modeled mainly on the basis of the utility maximization paradigm from rational decision theory, in which a traveler chooses the alternative that provides the highest utility among all available alternatives. However, the specific values of the utilities are not known to the researcher with certainty, and therefore they are treated as random variables (Ben-Akiva and Lerman 1985). Models developed by this method are called Random Utility Maximization (RUM) models.

Travel-behavior models based on the RUM paradigm were developed and applied in several different fields, including market research and social sciences. Specifically, in the area of travel demand, models were developed for mode choice, destination choice, land-use and transport interaction, and route choice. Modeling was also executed for systems requiring multi-dimensional decisions. The new approach of travel-demand models developed in the past decade; the activity-based model (ABM) approach is a representative example of multi-dimensional choice models.

ABM simplifies the simultaneity of many travel choices by grouping them into a hierarchical, multi-dimensional decision structure. This structure can be pictured as a Nested Multinomial Logit (Bowman and Ben-Akiva 1996). The nested model structure can be thought of as representing a hierarchical order of choices; in multi-dimensional choice problems, it is usually proposed according to logical order and experience, and it is usually fixed: i.e., the researcher assumes a single-choice, multi-dimensional hierarchical order even though there are a number of possible options. The researcher may test different structures, but this is usually done either for the entire population or for pre-defined segments. For example, assume that an ABM is composed of five choice decisions; e.g., activity purpose, departure time, mode, destination and route choice. In this case, there are theoretically 120 (5!) possible hierarchal orders. Moreover, there is little evidence in the literature that the logical choice hierarchy order proposed here best fits a given data set. Although researchers were aware of this problem (e.g., Forinash and Koppelman 1993), they offered no proven procedure on how to select the best hierarchical structure.

The Portland activity-based model (Bowman et al. 1998), for example, contains five types of sub-models, organized in a hierarchical structure. At the highest level is the activity-pattern model, which determines the purpose of a person's primary and secondary activities during the day, and tour type. The level that follows next in descending order is the time-of-day model, which predicts the combination of departure times from home and from the primary activity. At the third level, in the middle of this hierarchical structure, a mode and destination model is applied for each tour. The two bottom levels are the work-based sub-tour model, which is a mode-destination-choice model for the sub-tour that leaves from the work place and returns there; and the intermediate-stop model, a location-choice model for an intermediate activity during a given half-tour.

Similar models to the Portland ABM that were subsequently developed, such as the San Francisco ABM (Jonnalagadda et al. 2001), the Florida ABM (Pendyala et al. 2004), and the Jakarta ABM (Yagi 2006), kept its basic structure and choice order, with some modifications and improvements. The Atlanta ABM (Bradley and Vovsha 2005), for example, proposed a joint choice of daily activity-pattern types for all household members that explicitly took into account the added group-wise utilities of joint participation in the same activity. On the other hand, the Tel-Aviv ABM (Shiftan and Ben-Akiva 2011) has a different model structure. It considers the choices relating to the main activity (time, destination and the main tour mode) to be higher hierarchical choices, and therefore, the choice of the tour structure would follow these models.

Both the Portland ABM (and similar model structures) and the Tel-Aviv ABM demonstrate that there may be different structures for an ABM. Most of these models also assume a fixed-choice order for the pre-segmented study population. It might happen, though, that different groups of people would best be represented by different model structures representing different choice hierarchical orders. Moreover, estimating travel-demand models using the entire or pre-segmented study population, as is often the case, represents the behavior of an "average" person; it does not indicate differences and variety as do latent class models, which are required to better understand a population's travel behavior (see Hess et al. 2011). Presenting "average" results contradicts the assumption that different groups of people have different choice preferences. Therefore, there is need to segment travelers into homogeneous groups according to more than one shared travel-behavior or other characteristic and estimate a choice model for each segment. The main limitation in most current practices is that the structures of these models is treated generically for the entire population although different groups can receive different model structures; i.e., different hierarchical orders.

The objective of this paper is to develop an integrated methodological framework for multi-dimensional discrete choice models that accounts for the assumption that different groups of people have different elasticities of demand; thus, different model structures may best describe their behavior.

Methodology

The idea behind FMS

The core of the development of an integral methodological framework, the Flexible Model Structure (FMS), enhances the application of multi-dimensional discrete choice models by developing an optimization algorithm that segment given data and searches for the best multi-dimensional model structure for each segment. The optimization algorithm is able to suggest a segmentation process that obtains maximum homogeneity within segments and maximum heterogeneity among segments, and it searches for the best model structure in each segment.

The philosophy behind the FMS concept is illustrated in Fig. 1. Each individual may belong to any segment with a certain degree of membership (probability), according to *fuzzy segmentation theory* (Wedel and Kamakura 2000). These segments are latent, because their centroids are not directly observed as in the a priori segmentation process, but are inferred from observable variables that can be directly measured through average overall partial memberships. According to the same concept, each segment can belong to every multi-dimensional model structure with varying degrees of membership (probability).

The degree of flexibility of the FMS is compared to the customary approach, shown on the right side of Fig. 1. The customary approach can be viewed as a private case of FMS, in which either all individuals belong to one and only one segment (example a) or each individual belongs to only one of the pre-defined segments (example b—probability of 100 %); each segment, furthermore, has the same model structure as currently implemented in most multi-dimensional choice models, such as ABM.

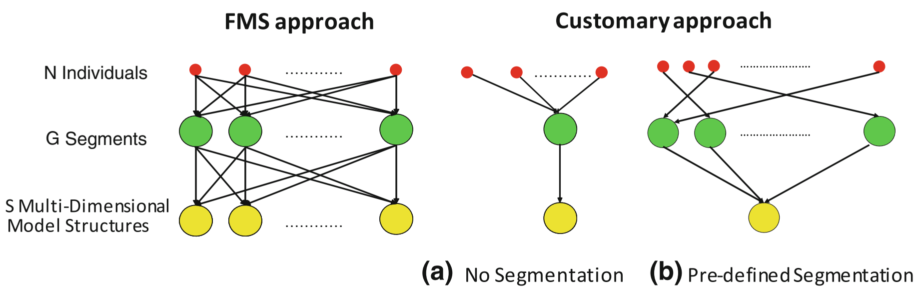


Fig. 1 Graphical representations of the FMS framework and the customary approach

Mathematical formulation of the FMS

Following the concept of the FMS, the model is composed of three parts: the alternative choice given the model structure, the structure choice given the segment and the segmentation process. This formulation can be represented mathematically as follows:

$$P_{ni} = \sum_{s=1}^{S} \sum_{g=1}^{G} P_n(i|s) \times P(s|g) \times P_n(g|x, \alpha_x) \tag{1}$$

where P_{ni} is the probability that individual n (n = 1,2,3…,N) will choose alternative i (i = 1,2,…I); $P_n(i|s)$ probability that individual n (n = 1,2,3…,N) will choose alternative i (i = 1,2,…I), given the model structure s (s = 1,2,…,S); $P(s|g)$ probability that a given segment g will choose structure s; in other words, the degree of membership of segment s in structure s; $P_n(g|x, \alpha_x)$ probability that individual n will choose segment g, given the individual's characteristics x and their weight, α_x, in the segmentation process; in other words, the degree of membership of individual n in segment g.

The first part of the FMS formulation is the alternative choice, given the model structure $P_n(i|s)$. This can be any discrete choice model having any model structure for choosing alternative i (i = 1,2,…I), given the model structure s (s = 1,2,…,S).

The second part of the FMS formulation is the probability that a given segment g will choose structure s; in other words, the degree of membership of segment g in structure s. This model can be described by any discrete choice model, $P_n(i|s)$; given that the Mixed Multinomial Logit (MMNL) (McFadden and Train 2000) is considered to be the most flexible discrete choice model, it is used here. An MMNL model is estimated for each segment, with the number of alternatives equal to the number of available model structures.

The third part of the FMS formulation is the segmentation process. In general, various segmentation processes can be represented here. In our case, the segmentation process deals with the degree of membership $P_n(g|x, \alpha_x)$, and it is based on the fuzzy method of segmentation, using the C-means algorithm (Dunn 1974; Bezdek 1974).

The degree of membership u_{ij} is calculated as follows:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{2}$$

where m is the any real number greater than 1 (m is usually 2); u_{ij} is the degree of membership of x_i in cluster j; $\left\| x_i^{(j)} - c_j \right\|$ is the distance measured between data point x_i and the segment center c_j; and C is the number of segments.

In the segmentation process, all variables are normally treated as if they have the same weight in determining how an object belongs to a segment. In practical terms, some variables might have greater influence in determining the membership of an individual to one particular segment than to others. This determination can be calculated by weighting single attributes for each segment as shown in Equation (3) (Keller and Klawonn 2000).

The distance between a datum x_k and a segment (vector) v_i is defined by

$$d^2(v_i, x_k) = \sum_{s=1}^{p} \alpha_{is}^{t} \cdot \left(x_k^{(s)} - v_i^{(s)}\right)^2 \tag{3}$$

where $x_k^{(s)}$ and $v_i^{(s)}$ is the s^{th} coordinates of the vectors x_k and v_i, respectively; I is the number of variables or attributes; α_{is}^{t} is the a parameter determining the influence of attribute s for segment i (according to Keller and Klawonn (2000) t is usually equal to 1).

In this study, the parameter α_{is} can take any number between 0 (no influence) and 1 (full influence). The distance ($d^2(v_i, x_k)$) as presented in Eq. (3) can take different shapes. Here, the distance takes a linear shape as presented in the following formula:

$$\text{Linear shape: } d_{ij}^{*} = a + b \cdot d_{ij} \tag{4}$$

where a, b are parameters to be estimated.

Example of a Detailed Mathematical Formulation of FMS

A simple example is presented below to demonstrate the mathematical formulation of FMS:

Model structure $P_n(i|s)$: Two available Nested Logit (NL) model structures represent the choice problem for individual n regarding a certain alternative i.
Structure choice $P(s|g)$: The MMNL model describes the probability of segment g choosing structure s.
Degree of membership $P_n(g|x,\alpha_x)$: The degree of membership of individual n in segment g is represented by fuzzy segmentation with weighting of data variables and a linear shape.

In this case, the probability that individual n will choose alternative i is presented as follows:

$$P_{ni} = \sum_{s=1}^{S}\sum_{g=1}^{G} P_n(i|s) \times P(s|g) \times P_n(g|x,\alpha_x) \tag{5}$$

$$= \sum_{s=1}^{S}\sum_{g=1}^{G}\left(\frac{e^{\frac{V_{is}}{\lambda_{ks}}}\left(\sum_{j\in B_k} e^{\frac{V_{js}}{\lambda_{ls}}}\right)^{\lambda_{ks}-1}}{\sum_{l=1}^{K}\left(\sum_{i\in B} e^{\frac{V_{js}}{\lambda_{ls}}}\right)^{\lambda_{ls}}}\right) \times \left(\frac{1}{R}\sum_{r=1}^{R}\left(\frac{e^{\beta' x_{sg}}}{\sum_j e^{\beta' x_{sg}}}\right)\right) \times \left(\frac{1}{\sum_{g=1}^{G}\left(\frac{x_i - c_{g'}}{x_i - c_g}\right)^2}\right)$$

$$x_i - c_g^2 = d^2\left(v_g, x_{wn}\right) = \sum_{d=1}^{D} \alpha_{dg} \cdot \left(x_{dn}^{(d)} - v_g^{(d)}\right)^2 \tag{6}$$

where i,1,…,I—number of alternatives; n,1,…,N—number of individuals; s,1,…,S—number of model structures; g,1,…,G—number of segments; r,1,…,R—number of draws in the

MMNL; V_{js}—observed utility of alternative j of model structure s; λ_{ks}—nest coefficient of nest k for model structure s; $d^2(v_g, x_{wn})$—D-dimension square-distance between observation n and segment g; α_{dg}—weight of variable d belonging to segment g; $\beta' x_{sg}$—utility of segment g of model structure s, with different values of β' drawn from the normal distribution.

Empirical implementation

The empirical implementation uses the Tel-Aviv tour-based data source (Cambridge Systematics, 2008). Three different models were tested in order to conceptualize a multi-dimensional choice model with two choices, destination and mode, two of the models being nested structures: (1) Nested Logit—Destination-Mode (NL_DM); (2) Nested Logit—Mode-Destination (NL_MD); (3) FMS. The data for the first two models, NL_DM and NL_MD, were segmented into two: segment 1, trips for work purposes; segment 2, trips for non-work purposes. This segmentation represents the common practice in such models.

In order to simplify and demonstrate the idea of estimating FMS, data on 20,000 tours that were carried out as the main destination in the city of Tel-Aviv (divided into 9 super zones) and by means of three main travel modes—car as a driver, car as a passenger, and bus—were selected. To further simplify the estimation exercise, a select, limited set of variables was used:

- *Gender*: dummy variable, which takes the values of 1 for male and 0 for female.
- *Number of Cars per Household*: ordered categorical variable, which takes the values of 0, 1, 2, and 3, the last for three or more cars per households.
- *Number of Persons per Household*: ordered categorical variable, which takes the values of 1–8, the latter for eight or more persons per household.
- *Car and Bus Travel Time*: continuous variables, which describe the total door-to-door travel time from home to main destination, in minutes.

For the Nested Logit models, a further choice of additional explanatory variables was used to show the advantages of FMS over the traditional approach:

- *Tour Purpose*: dummy variable, which takes the values of 1 if the tour purpose is work, and 0 if otherwise; this variable is used to segment the data.
- *Age*: dummy variable, which takes the values of 1 if the person is between the ages of 30 and 60, and 0 if otherwise.
- *Tour Complexity*: dummy variable, which takes the value of 1 when there is at least one intermediate stop before or/and after the primary destination, and 0 if otherwise (i.e., a simple tour from home to the main destination and back, with no other stops).
- *Work Status*: dummy variable, which takes the values of 1 if the person works (whether full or part time), and 0 if otherwise.
- *Household Workers*: ordered categorical variable, which describes the number of workers in the household.
- *Number of Licenses*: ordered categorical variable, which describes the number of persons holding a driving license.

Basic analyses of the data presented in Table 1 show that 61 % of the tours were made by car as a driver, 7 % by car as a passenger, and 32 % by bus. The data for the first two models were segmented into two segments: 33 % of the tours belong to segment 1, work purpose, and 67 % to segment 2, non-work purpose.

Table 1 Tour data analyses

Variable	Value	Frequency (%)
Household size	1	21
	2	27
	3	18
	4	17
	5	11
	6	4
	7	1
	8+	1
Number of cars in household	0	13
	1	44
	2	38
	3+	5
Gender	Men	49
	Women	51
Number of workers in household	0	19
	1	35
	2	38
	3+	8
Age	Between 30 and 60	53
	Other	47
Work status	Work	72
	Does not work	28
Licensed drivers in household	0	2
	1	31
	2	47
	3+	20
Tour complexity	Not complex	66
	Complex	34
Trip purpose	Work	33
	Non-work	67
Mode choice	Driver	61
	Passenger	7
	Bus	32

The first two models, which employ a pre-segmented Nested Logit structure, illustrate the implementation of a fixed model structure for pre-segmented data as is common in multi-dimensional choice models. The third model implements the FMS to demonstrate that more than one model structure can fit the data. The first model, a two-segment Nested Logit in which Destination Choice is on the higher level and Mode Choice on the lower (NL_DM), assumes that different mode alternatives for one destination share common, unobserved errors. The utility functions of the 27 alternatives in this model (9 destinations and 3 modes), which contain 21 parameters ($\beta_1,\ldots,\beta_{21}$) for each segment, are described by

Eqs. (7–9) for the three modes. In addition, there are 9 parameters $(\lambda_1, \ldots, \lambda_9)$ representing 9 super-zone nest coefficients:

$$
\begin{aligned}
V_{\mathrm{Driver,SZ1-9}} = {} & \upbeta_1 + \upbeta_2 \cdot \mathrm{Gender} + \upbeta_3 \cdot \mathrm{Ncars} + \upbeta_4 \cdot \mathrm{HHsize} + \upbeta_5 \cdot \mathrm{Age} + \upbeta_6 \cdot \mathrm{Complex} \\
& + \upbeta_7 \cdot \mathrm{WorkStatues} + \upbeta_8 \cdot \mathrm{HHworkers} + \upbeta_0 \cdot \mathrm{NumLicrnse} \\
& + \upbeta_{10} \cdot \mathrm{CarTime}_{\mathrm{SZi}}
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
V_{\mathrm{Passenger,SZ1-9}} = {} & \upbeta_{11} + \upbeta_{12} \cdot \mathrm{Gender} + \upbeta_{13} \cdot \mathrm{Ncars} + \upbeta_{14} \cdot \mathrm{HHsize} \\
& + \upbeta_{15} \cdot \mathrm{Age} + \upbeta_{16} \cdot \mathrm{Complex} \\
& + \upbeta_{17} \cdot \mathrm{WorkStatues} + \upbeta_{18} \cdot \mathrm{HHworkers} \\
& + \upbeta_{19} \cdot \mathrm{NumLicrnse} + \upbeta_{20} \cdot \mathrm{CarTime}_{\mathrm{SZi}}
\end{aligned} \tag{8}
$$

$$
V_{\mathrm{Bus,SZ1-9}} = \upbeta_{21} \cdot \mathrm{BusTime} \tag{9}
$$

The second model is also a two-segment Nested Logit in which Mode Choice is on the higher level and Destination choice on the lower (NL_MD); it assumes that the different destination alternatives for each mode share common, unobserved errors. The utility functions are similar to those in the previous model, containing the same 21 parameters $(\beta_1, \ldots., \beta_{21})$ for each segment. In addition, there are 3 parameters$(\lambda_1, \lambda_2, \lambda_3)$ representing 3 mode nest coefficients. The estimation process for the NL models was carried out using BIOGEME software (Bierlaire et al. 2008), and that for FMS through MATLAB software (Matworks 2000), since the mathematical representation was new.

The results of the estimated parameters for the Nested Logit, Destination-Mode, are presented in Table 2. Most of the parameters are statistically significant at the 5 % level (see *t* test column). The estimation results have a final likelihood value of −52,352.4 (−17,113.1 for segment 1 and −35.239.3 for segment 2), and an adjusted ρ^2 of 0.191.

The results of the estimated parameters for the Nested Logit, Mode-Destination, are presented in Table 3. As in the Destination-Mode case, most of the parameters are statistically significant at the 5 % level (see *t* test column). The final likelihood value is −52,605.1 (−17,157.5 for segment 1 and −35,447.6 for segment 2), and the adjusted ρ^2 is 0.187. A comparison of the two estimation results shows that the NL Destination-Mode structure is preferable to that of the NL Mode-Destination (a final likelihood of −52,352.4 with 60 parameters, 30 for each segment, compared with a likelihood of −52,605.1 with 48 parameters, 24 for each segment). As mentioned before, the implementation of these two models for a multi-dimensional choice model illustrates a fixed model structure for pre-segmented data as reported in the literature.

As discussed above, the FMS model is composed of three parts: the alternative choice models, the structure choice model, and the segmentation process. The application for our case is shown in Fig. 2. Again, as has been mentioned, for simplicity sake, only a few main explanatory variables are included in this model. This procedure can actually better demonstrate the power of the FMS when comparing it to the NL model with more explanatory variables. The alternative choice model has the same utility functions of 27 alternatives (9 destinations and 3 modes), with only 5 variables (Gender, N cars, HH size, Car Time, Bus Time), compared to 11 variables in the Nested Logit models (the 5 used in the FMS model plus Age, Complexity, Work Status, HH workers, Num Licenses and Trip Purpose) that are used to segment the data); and 11 parameters $(\beta_1,\ldots.,\beta_{11})$, compared to 21 for each segment in the NL models. The utility functions are described by Eqs. 10–12 for the three modes. In addition, there are 9 parameters $(\lambda_1, \ldots, \lambda_9)$ representing 9 super-zone nest coefficients for

Table 2 NL_DM estimated utility parameters

Mode	Variable	Segment 1—work trips	Segment 2—nonwork trips
Driver	Constant	−1.68 (−12.5)	−1.94 (−22.4)
	Gender	0.77 (11.7)	0.84 (20.0)
	N cars	1.35 (22.6)	1.15 (31.5)
	HH size	0.11 (2.7)	0.08 (2.9)
	Age	0.20 (2.9)	0.12 (3.0)
	Tour complex	0.27 (3.8)	0.17 (3.9)
	Work status	0.16 (*1.5*)	0.06 (*0.9*)
	HH workers	−0.15 (−2.6)	−0.08 (−2.2)
	Num licenses	−0.12 (*−1.8*)	−0.09 (−2.0)
	Travel time	−0.066 (−35.9)	−0.067 (−46.6)
Passenger	Constant	−1.99 (−10.4)	−1.74 (−17.7)
	Gender	−0.35 (−2.8)	−0.23 (−3.6)
	N cars	0.48 (5.3)	0.45 (9.9)
	HH size	0.07 (*1.1*)	0.165 (5.1)
	Age	−0.12 (*−0.9*)	−0.19 (−3.2)
	Tour complex	−0.42 (−2.8)	0.06 (*0.93*)
	Work status	−1.04 (−6.4)	−1.26 (−16.0)
	HH workers	0.03 (*0.27*)	0.05 (*1.6*)
	Num licenses	−0.21 (−2.1)	−0.24 (−4.9)
	Travel time	−0.048 (−8.4)	−0.060 (−20.7)
Bus	Travel time	−0.031 (−24.5)	−0.031 (−42.8)
Superzone 1	Nest coefficient	0.730(5.2)	0.714(8.0)
Superzone 2	Nest coefficient	0.704(5.65)	0.568(10.3)
Superzone 3	Nest coefficient	1.00 (18.6)	1.00 (12.2)
Superzone 4	Nest coefficient	0.935(*1.4*)	0.926 (2.3)
Superzone 5	Nest coefficient	1.00 (12.5)	1.00 (23.8)
Superzone 6	Nest coefficient	1.00(22.3)	0.962(*1.2*)
Superzone 7	Nest coefficient	0.758 (4.7)	0.671 (8.3)
Superzone 8	Nest coefficient	0.735(5.3)	0.637(10.0)
Superzone 9	Nest coefficient	0.782 (4.2)	0.730 (7.3)
Observations		6712	13288
Initial likelihood		−21850	−42917.2
Final likelihood		−17113.1	−35239.3
Number of parameters		30	30
Adjusted rho-bar squared		0.215	0.178

Note *T* test (italicised *T* test): : statistically insignificant at the 5 % significance level

the Nested Logit destination-mode model, and 3 parameters $(\lambda_1, \lambda_2, \lambda_3)$ representing 3 mode nest coefficients for the Nested Logit mode-destination model.

$$V_{Driver,SZ1-9} = \beta_1 + \beta_2 \cdot Gender + \beta_3 \cdot Ncars + \beta_4 \cdot HHsize + \beta_5 \cdot CarTime_{SZi} \quad (10)$$

$$V_{Passenger,SZ1-9} = \beta_6 + \beta_7 \cdot Gender + \beta_8 \cdot Ncars + \beta_9 \cdot HHsize + \beta_{10} \cdot CarTime_{SZi} \quad (11)$$

$$V_{Bus,SZ1-9} = \beta_{11} \cdot BusTime \quad (12)$$

Table 3 NL_MD Estimated Utility Parameters

Mode	Variable	Segment 1—work trips	Segment 2—nonwork trips
Driver	Constant	−1.29 (−6.9)	−1.72 (−13.2)
	Gender	0.88 (12.0)	1.01 (20.8)
	N cars	1.56 (24.4)	1.39 (36.0)
	HH size	0.07 (*1.3*)	0.05 (*1.6*)
	Age	0.24 (3.1)	0.16 (3.2)
	Tour complex	0.31 (3.8)	0.21 (4.1)
	Work status	0.23 (*1.9*)	0.10 (*1.3*)
	HH workers	−0.20 (−2.8)	−0.12 (−2.6)
	Num licenses	−0.05 (*−0.6*)	−0.005 (*−0.1*)
	Travel time	−0.011 (*−1.2*)	−0.012 (−1.98)
Passenger	Constant	−1.31 (−3.4)	−1.33 (−13.8)
	Gender	−0.44 (−3.0)	−0.31 (−4.1)
	N cars	0.56 (5.5)	0.55 (10.0)
	HH size	0.03 (*0.4*)	0.15 (3.8)
	Age	−0.16 (*−1.1*)	−0.24 (−3.3)
	Tour complex	−0.49 (−2.9)	0.07 (*0.94*)
	Work status	−1.16 (−6.2)	−1.49 (−4.8)
	HH workers	−0.02 (*−0.1*)	0.003 (*0.1*)
	Num licenses	−0.17 (*−1.5*)	−0.17 (−2.9)
	Travel time	−0.008 (*−0.6*)	−0.008 (−16.1)
Bus	Travel time	−0.013 (−2.4)	−0.009 (−3.1)
Driver	Nest coefficient	0.169 (2.2)	0.189(2.2)
Passenger	Nest coefficient	0.154 (2.0)	0.154 (8.9)
Bus	Nest coefficient	0.382 (2.4)	0.243 (2.4)
Observations		6712	13288
Initial likelihood		−21850	−42917.2
Final likelihood		−17,157.50	−35,447.60
Number of parameters		24	24
Adjusted rho-bar squared		0.214	0.173

Note *T* test (italicised *T* test): statistically insignificant at the 5 % significance level

The segmentation process includes a fuzzy method (C-mean algorithm) with weighting variables, which indicates the influence of each variable in determining the membership of an object in a segment. Moreover, the distance of an object to a segment takes a linear shape as in Eq. 4. Following the "Gap Statistics" method (Tibshirani et al. 2001), the data were divided into three segments.

The alternative choice model includes two possible model structures: Nested Logit Destination-Mode and Nested Logit Mode-Destination, representing a combination of Model 1 and Model 2. The structure choice implemented in this model is composed of three Mixed Logit models, one for each segment, in which the utility function of a segment choosing structure 1 (NL_DM) is V_{NL1}, and that of a segment choosing structure 2 (NL_MD) is V_{NL2}, which is set at zero:

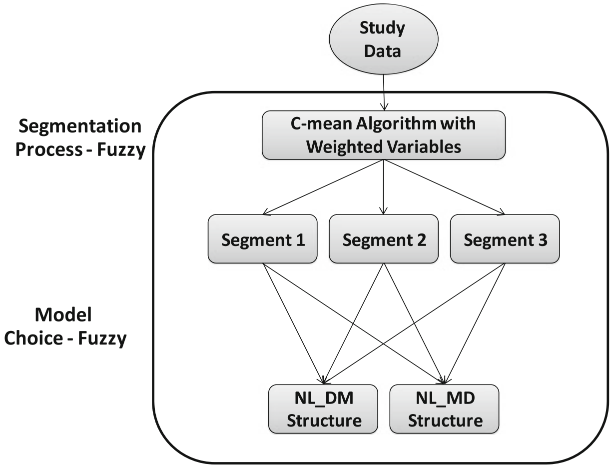


Fig. 2 FMS model structure

$$V_{NL1} = \beta_1 + (\beta_2 + \sigma_2) \cdot Gender + (\beta_3 + \sigma_3) \cdot Ncars + \beta_4 \cdot HHsize + \beta_5 \cdot CareTime + \beta_6 \cdot BusTime \quad (13)$$

$$V_{NL2} = 0 \quad (14)$$

where σ_2 is the Standard Deviation *of normal distribution with a mean of* β_2; σ_3 is the Standard Deviation *of normal distribution with a mean of* β_3.

It is important to mention that in contrast to the mode or destination choice models, time coefficients in the structure choice model can receive a positive value. A positive time coefficient indicates that as travel time increases, the structure choice model, whose utility contains the time coefficient, is preferable to the other models, and that is logical.

In this model, therefore, we estimate 9 parameters for the characteristics of the segments, 9 parameters for the weights of the variables, 2 parameters for the linear distance, 24 parameters for the structure choice model, 60 (3*20) parameters for the NL_DM models and 42 (3*14) parameters for the NL_MD models. In all, there are 146 parameters to be estimated. As with the previous models, the characteristics of the segments are calculated as a weighted average of the variables, taking into account the degree of membership of an object in a segment, which is produced from the estimation process. According to Table 4,

Table 4 Segment center characteristics

Variable	Segment 1	Segment 2	Segment 3
Gender	59 % men, 41 % women	33 % men, 67 % women	41 % men, 59 % women
Number of cars per HH	1.6	1.5	0.1
Household size	3.2	2.6	2.0
Actual mode choice	72 % D, 6 % P, 22 % B	66 % D, 8 % P, 26 % B	5 % D, 9 % P, 86 % B
Average membership degree	58 %	28 %	14 %

D Driver, *P* Passenger, *B* Bus (average choice: 61, 7 and 32 %, respectively)

Table 5 Segment variable weights

Variable	Segment 1	Segment 2	Segment 3
Gender	1.0	0.94	0.06
Number of cars per HH	0.19	0.04	0.79
Household size	0.20	0.35	0.01
$d_{ij}^{*} = 0.115 - 0.210 \cdot d_{ij}$			

each segment is characterized and tagged by three variable values and by actual mode choice. Overall, 58 % of the study population belongs to segment 1, 28 % to segment 2 and 14 % to segment 3. The weights of the influence of the variables on a segment's membership are presented in Table 5. For segments 1 and 2 the variable that most influences segment membership is "Gender"; and for segment 3, "Number of cars."

The results of model structure 1 (NL_DM) and model structure 2 (NL_MD) for each segment are presented in Table 6 and Table 7, respectively. These results show that segment 1 in both models, which is characterized as Mixed Gender with 1.6 cars per household and an average household size of 3.2 persons, can be tagged the "Drivers" segment. In the first model structure, the constant, gender and number of cars coefficients of the "Driver" alternative for segment 1 receive high positive values, where as low negative value characterize the coefficients of household size and car travel time; this compares with negative signs for travel time for "Passenger" and "Bus." Segment 2 in

Table 6 Estimated utility parameters—model structure 1—NL_DM

Mode	Variable	Segment 1	Segment 2	Segment 3
Driver	Constant	1.39 (4.4)	−0.81 (−8.5)	−6.99 (−19.9)
	Gender	2.22 (15.3)	0.27 (5.9)	2.11 (9.3)
	No. of cars	1.00 (5.2)	0.075 (2.1)	4.98 (28.3)
	Household size	−0.14 (−2.1)	−0.17 (−10.3)	0.30 (4.6)
	Travel time	−0.055 (−37.3)	−0.047 (−18.1)	−0.132 (−18.2)
Passenger	Constant	−0.33 (*−1.6*)	−3.17 (−14.5)	−1.85 (−10.9)
	Gender	−0.21 (*−1.9*)	−0.86 (−4.3)	0.19 (1.98)
	No. of cars	−0.01 (*−0.05*)	0.52 (5.6)	−0.33 (−4.7)
	Household size	−0.05 (*−1.42*)	−0.26 (−6.0)	0.003 (*0.08*)
	Travel time	−0.039 (−9.6)	−0.045 (−4.9)	−0.065 (−11.1)
Bus	Travel time	−0.022 (−8.5)	−0.029 (−25.8)	−0.027 (−15.3)
Superzone 1	Nest coefficient	0.251 (2.6)	0.321 (6.4)	0.476 (6.5)
Superzone 2	Nest coefficient	0.243 (2.3)	0.370 (9.0)	0.352 (5.3)
Superzone 3	Nest coefficient	1.00 (3.9)	1.00 (4.9)	0.936 (8.1)
Superzone 4	Nest coefficient	0.672 (5.9)	0.948 (15.1)	0.738 (8.1)
Superzone 5	Nest coefficient	1.00 (5.9)	1.00 (6.3)	1.00 (6.6)
Superzone 6	Nest coefficient	0.709 (5.8)	1.00 (10.1)	1.00 (9.6)
Superzone 7	Nest coefficient	0.400 (4.1)	0.465 (8.9)	0.584 (7.1)
Superzone 8	Nest coefficient	0.385 (5.5)	0.248 (*1.7*)	0.745 (9.8)
Superzone 9	Nest coefficient	0.570 (6.3)	0.576 (9.9)	0.637 (7.5)

Note T test (italicised *T* test)—insignificant at the 5 % significance level

Table 7 Estimated utility parameters—model structure 2—NL_MD

Mode	Variable	Segment 1	Segment 2	Segment 3
Driver	Constant	3.56 (5.1)	3.46 (9.1)	−7.01 (−7.3)
	Gender	4.55 (13.3)	−2.99 (−2.8)	0.45 (*0.82*)
	No. of cars	−0.36 (*−1.6*)	−0.48 (*−1.8*)	4.87 (14.7)
	Household size	−0.34 (−2.4)	−1.69 (−10.2)	−2.22 (−7.6)
	Travel time	−0.071 (−4.9)	−0.048 (*−0.9*)	−0.124 (−7.3)
Passenger	Constant	−1.23 (−2.0)	3.20 (4.5)	−5.16 (−3.2)
	Gender	−0.73 (−2.4)	2.29 (3.7)	−4.91 (−4.7)
	No. of cars	0.13 (*0.5*)	−3.91 (−8.6)	−6.29 (−16.2)
	Household size	−0.12 (*−1.0*)	0.90 (3.0)	−0.91 (−5.2)
	Travel time	−0.035 (*−1.0*)	−0.048 (*−1.8*)	−0.001 (*0*)
Bus	Travel time	−0.097 (−3.6)	−0.052 (−6.1)	−0.026 (−5.4)
Driver	Nest coefficient	0.467 (4.9)	0.247 (*0.74*)	0.758 (2.1)
Passenger	Nest coefficient	0.248 (*0.94*)	0.457 (*1.87*)	0.456 (*0.4*)
Bus	Nest coefficient	1.00 (*1.92*)	0.727 (3.7)	0.564 (4.7)

Note *T* test (italicised *T* test)—insignificant at the 5 % significance level

both model structures can be tagged as "Mixed Modes," and segment 3 as "Bus Riders"; segment 3 is very sensitive to car travel time (−0.132 for model structure 1 and −0.124 for model structure 2).

The structure-choice parameter results, as presented in Table 8, show that the time coefficients in segment 1 are positive (significant at the 10 % level). That is to say, as travel times increase, it is more likely that structure 1 will be chosen (NL_DM). If the value of the characteristics of each segment is taken into account, the probability of choosing model structure 1 (NL_DM) and model structure 2 (NL_MD) for segment 1 is 78 and 22 %, respectively. The probability of choosing model structure 1 (NL_DM) for segment 2 is almost 100 %; on the other hand, the probability of choosing model

Table 8 FMS estimated utility parameters—structure choice

Variable	Segment 1		Segment 2		Segment 3	
	NL_DM	NL_MD	NL_DM	NL_MD	NL_DM	NL_MD
Constant	−6.37 (−6.6)	–	1.56 (*1.7*)	–	−3.49 (−3.0)	–
Gender	0.20 (*1.26*)	–	4.33 (14.4)	–	−0.39 (−2.0)	–
No. of cars	−0.152 (*−1.0*)	–	1.84 (7.0)	–	−2.78 (−18.4)	–
Household size	0.27 (4.2)	–	0.42 (3.4)	–	0.06 (*0.7*)	–
Car travel time	0.130 (*1.8*)	–	−0.132 (*−1.2*)	–	0.140 (*1.5*)	–
Bus travel time	0.064 (*1.9*)	–	0.026 (*0.63*)	–	0.028 (*0.7*)	–
Sigma_Gender	±0.002 (*0.1*)	–	±1.68 (11.1)	–	±0.247 (*1.7*)	–
Observations		20,000				
Initial likelihood		−64,767.2				
Final likelihood		−50,852.7				
Number of parameters		146				
Adjusted rho-bar squared		0.213				

Note *T* test (italicised *T* test)—insignificant at the 5 % significance level

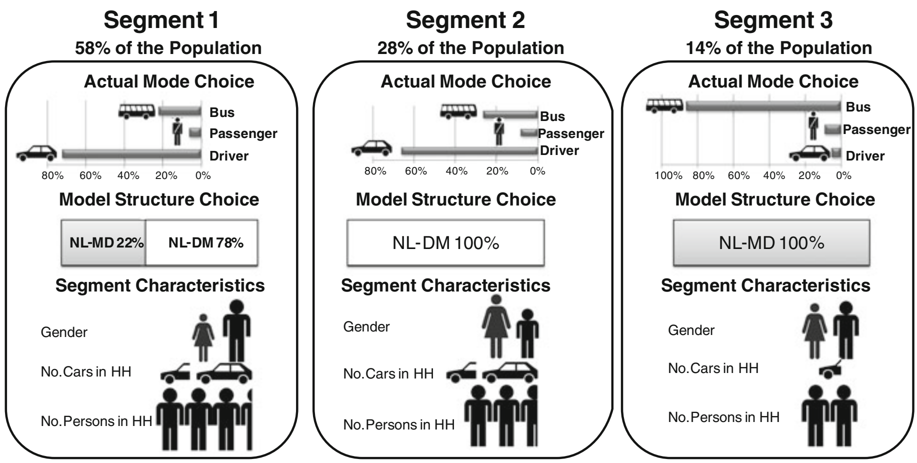


Fig. 3 The travel-demand story

structure 1 for segment 3 is almost nil. These results show that the behavior of segments that are more likely to choose the "Driver" alternative for their trips (segments 1 and 2) is better described by model structure 1 (NL_DM). It is remarkable that also various standard deviation coefficients were tested the only significant one at the 5 % level was for "Gender" in segment 2. Therefore, the structure choice model for segments 1 and 3 is reduced to a simple Multinomial Logit.

The results given in Tables 4 through 8 are illustrated in Fig. 3, which summarizes the three dimensions of travel-demand with the given data: segment characteristics, segment model structure, and segment actual choice. The estimation results, presented in Table 8, show an improvement in the final likelihood value compared to the previous two models (−50,852.7 for FMS with 146 parameters, compared with −52,352.4 for NL_DM with 60 parameters and −52,605.1 for NL_DM with 48 parameters). The adjusted ρ^2 indicates that the improvement is significant despite the additional 114 parameters in FMS (146 for FMS, compared with only 60 for NL_DM): 0.213 compared with 0.191, respectively.

Discussion and Conclusions

This study developed an integral methodological framework, the FMS, to enhance the application of the multi-dimensional discrete choice model through the development of an optimization algorithm that segments given data and searches for the best model structure for each segment. The optimization algorithm is able to suggest a segmentation process that obtains maximum homogeneity within groups and maximum heterogeneity between groups, and searches for the best model structure match for each group.

The segmentation process, which relies on the fuzzy method, derives from the assumption that an individual belongs to each segment in varying degrees of membership. Moreover, the variables used in the segmentation process are treated with different weights in determining the degree to which an individual belongs to a segment. The partial membership of each individual in a segment determines the characteristics of that segment. These segments are latent classes, since their characteristics are not directly observed but, rather, are inferred from other variables that are observed and directly measured. The model structure search relies on the same concept, with each segment belonging to a certain degree of membership to every model structure.

Three models conceptualizing a multi-dimensional choice model of mode and destination as part of an Activity Based Model were tested. The first two models represented the estimation of a fixed-structure model, Nested Logit, for pre-segmented data, as is common in the implementation of multi-dimensional choice models. The third model, FMS, includes a fuzzy segmentation method with weighted variables and a combination of two or more model structures. The integration of a segmentation process and more than one model structure, as in FMS, significantly improves the estimation results over the fixed-model structure, as in first two models, though with fewer explanatory variables. The results support the hypotheses that claim that different model structures may best describe the behavior of different groups of people in multi-dimensional choice models. The implementation of FMS presents the travel behavior of an individual as a mix of travel behaviors, represented by a number of segments, and the choice model of each segment as a combination of different choice model structures. The FMS model breaks the consensus that an individual should belong to only one segment and that a segment can take only one structure.

The implementation of FMS in this paper demonstrated the approach of the segmentation process and the model structure search for a two-dimensional discrete choice model. This approach can make a significant contribution to the development and estimation of travel-demand models involving more than one choice dimension and activity-based models in particular. However, an activity-based model system should try to estimate much more than two-dimensional travel choices; estimates should include the activity type, its timing and duration, destination, additional destinations in the tour and their details, and mode. Obviously the suggested approach can be of more importance to problems involving more than two-choice dimensions; at the same time, though, the complexity of this approach will also increase. The objective of this paper was to introduce the FMS approach and to test it on a simple model. As computation power continuously increases, it should be possible to extend the FMS approach to more complex problems with more choice dimensions and explanatory variables; however, future research is needed to generalize its use among full Activity Based Models.

Acknowledgments The authors would also like to express their appreciation for the contribution of three anonymous reviewers.

References

Ben-Akiva, M.E., Lerman, S.: Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambrige, MA (1985)

Bezdek, J.C.: Numerical taxonomy with fuzzy sets. J. Math. Biol. **1**, 57–71 (1974)

Bierlaire, M.: An Introduction to BIOGEME Version 1.6, biogeme.epfl.ch (2008)

Bowman, J., Ben-Akiva, M. E.: Activity based travel forecasting. New Orleans Conference (1996)

Bowman, J., Bradley, M., Shiftan, Y., Lawton, T.K., Ben-Akiva, M.E.: Demonstration of an activity based model system for Portland. Eighth World Conference on Transport Research, Antwerp (1998)

Bradley, M., Vovsha, P.: A model for joint choice of daily activity pattern types of household members. Transportation **32**(5), 545–571 (2005)

Cambridge Systematics: Tel-Aviv activity schedule model: model application. Report prepared for the Israel Ministry of Transport (2008)

Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in determining compact well-separated clusters. J. Cybern. **3**(3), 32–57 (1974)
Forinash, C., Koppelman, F.S.: Application and interpretation of nested logit models of intercity mode choice. Transp. Res. Rec. **1413**, 98–106 (1993)
Hess, H., Stathopoulos, A., Daly, A.: Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. Transportation **39**(3), 565–591 (2011)
Jonnalagadda, N., Freedman, J., Davidson, W., Hunt, J.D.: Development of a microsimulation activity-based model for San Francisco—destination and mode choice models. Transp. Res. Rec. **1777**, 25–35 (2001)
Keller, A., Klawonn, F.: Fuzzy clustering with weighting of data variables. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **8**(6), 735–746 (2000)
Matworks: MATLAB—The Language of Technical Computing. The MathWorks, Inc., Natick, MA (2000)
McFadden, D., Train, K.: Mixed MNL models for discrete response. J. Appl. Econom. **15**(5), 447–470 (2000)
Pendyala, R., Kitamura, R., Kikuchi, A.: FAMOS: The Florida activity mobility simulator. Progress in Activity-Based Analysis Conference, The Netherlands (2004)
Shiftan, Y., Ben-Akiva, M.E.: A practical policy-sensitive, activity-based, travel-demand model. Ann. Reg. Sci. **44**, 517–544 (2011)
Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. **63**(2), 411–423 (2001)
Wedel, M., Kamakura, W.: Market Segmentation: Conceptual Methodological Foundations, 2nd edn. Kluwer Academic Publishers, Boston (2000)
Yagi, S.: An activity-based microsimulation model of travel demand in the Jakarta Metropolitan Area. 11th International Conference on Travel Behaviour Research, Kyoto (2006)

Author Biographies

Dr. Robert Ishaq is Research Associate in the Transportation Research Institute, Faculty of Civil and Environmental Engineering at the Technion—Israel Institute of Technology. His research interests include travel behavior and demand models. He has extensive experience in transportation master plans, transit network planning, and traffic and transit assignments.

Shlomo Bekhor is Associate Professor in the Transportation and Geo-Information Department, Faculty of Civil and Environmental Engineering at the Technion—Israel Institute of Technology. His research interests include route choice modeling, large-scale equilibrium models, and innovative transport systems. He has extensive experience in transportation master plans, economic evaluations, and traffic and transit assignments.

Prof. Yoram Shiftan is the Head of the Transportation and Geo-Information Department in the Faculty of Civil and Environmental Engineering at the Technion, the Israel Institute of Technology. He is the editor of Transport Policy and the vice chair of the International Association of Travel Behavior Research (IATBR).