A LATENT-SEGMENTATION BASED APPROACH TO INVESTIGATING THE SPATIAL TRANSFERABILITY OF ACTIVITY-TRAVEL MODELS

Zeina Wafa
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712
Phone: 512-471-4535, Fax: 512-475-8744
Email: zww00@utexas.edu

Chandra R. Bhat (corresponding author)
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712
Phone: 512-471-4535, Fax: 512-475-8744
Email: bhat@mail.utexas.edu

Ram M. Pendyala
Georgia Institute of Technology
School of Civil and Environmental Engineering
Mason Building, 790 Atlantic Drive, Atlanta, GA 30332-0355
Tel: 404-894-2201; Fax: 404-894-2278
Email: rampendyala@gmail.com

Venu M. Garikapati
Arizona State University
School of Sustainable Engineering and the Built Environment
Tempe, AZ 85287-3005
Tel: (480) 965-3589; Fax: (480) 965-0557
Email: venu.garikapati@asu.edu

ABSTRACT
Spatial transferability of travel demand models has been an issue of considerable interest, particularly for small and medium sized planning areas that often do not have the resources and staff time to collect large scale travel survey data and estimate model components native to the region. With the advent of more sophisticated microsimulation-based activity-travel demand models, the interest in spatial transferability has surged in the recent past as smaller metropolitan planning organizations seek to take advantage of emerging modeling methods within the limited resources they can marshal. Traditional approaches to identifying geographical contexts that may borrow and transfer models between one another involve the exogenous *a priori* identification of a set of variables or criteria that are used to characterize the similarity between geographic regions. However, this *ad hoc* procedure presents considerable challenges as it is difficult to identify the most appropriate criteria a priori. To address this issue, this paper proposes a latent segmentation approach whereby the most appropriate criteria for identifying areas with similar profiles are determined endogenously within the model estimation phase. In other words, the relationships embedded in the data set help identify the optimal set of criteria that can be used to cluster regions according to their similarity with respect to activity-travel characteristics of interest. The methodology is demonstrated and its efficacy established through a case study in this paper that utilizes the National Household Travel Survey (NHTS) data set. It is found that the methodology offers a robust mechanism for identifying latent segments and establishing criteria for assessing transferability of models between areas.

Keywords: spatial transferability, activity-travel model, geographic contexts, MDCEV model, latent segmentation approach, regional similarity

1. INTRODUCTION

There is considerable interest among the transportation planning and modeling community in the notion of spatial transferability of travel demand models. Spatial transferability of a model refers to the ability to use a model that was estimated in one context in a different application context, and obtain useful results that simulate, more or less, locally observed behavior in the application context. While it is generally considered good practice to develop models based on locally collected data, there are situations where spatial transfer of models may be of interest. Some regions, particularly small and medium-sized planning organizations, may not have the resources and staff time necessary to undertake large scale survey data collection efforts and thus borrow models from other regions (Sikder et al, 2012). When such model transfer is considered, it is important to ensure that the transferred model offers useful and valid information in the application context (Koppelman and Wilmot, 1982).

There are two main approaches adopted in the literature for spatial transferability: application-based transferability and estimation-based transferability. In the context of the former, a model is developed and estimated based on data from one region, denoted as the *estimation context*. The model is then applied in the *application context* where its predictive abilities are assessed. Numerous studies in the literature adopt this approach to model transferability (e.g., Koppelman and Rose, 1983; Koppelman and Wilmot, 1986; Koppelman and Pas, 1986; Wilmot, 1995; Arentze *et al.*, 2002; Bekhor and Prato, 2009; Nowrouzian and Srinivasan, 2012; Sikder and Pinjari, 2013a; Sikder *et al.*, 2013b). On the other hand, in the method of estimation-based transferability assessment, also referred to as joint context estimation, the model is developed and estimated based on data from both contexts, the *estimation* and *application* areas. "Difference" parameters are included in the model to account for the fact that data comes from more than one source. Statistical tests are then carried out to determine whether the difference variables are significant, i.e., whether the parameters on a certain variable are essentially different between the two (or more) contexts. In this approach, common parameters can be estimated for variables for which data is limited (see Karasmaa 2001; Sikder et al., 2013b; Bowman et al., 2014). The estimation-based approach allows for statistical tests on the differences between coefficients and thus allows for a wide range of hypotheses. Another advantage is that it allows for the transferability assessment of parameters associated with specific variables, while the application-based transferability approach focuses exclusively on assessing transferability of the model as a whole.

The literature identifies a number of criteria or factors that may be considered when determining the nature of contexts for which transferability can be successfully achieved. The literature suggests similar transit service quality (McComb, 1986; Stopher et al., 2003; Mohammadian and Zhang, 2007), income levels (Caldwell and Demetsky, 1980; Wilmot, 1995; Reuscher et al., 2002), demographic characteristics (Caldwell and Demetsky, 1980; Mohammadian and Zhang, 2007), city socio-economic composition (McComb, 1986), and area size and type (McComb, 1986, Stopher et al., 2003; Reuscher et al., 2002; Everett, 2009) are significant determinants of the success of transferability of a model from one setting to another. Also, it has been found that intra-state transferability outperforms inter-state transferability (Sikder and Pinjari, 2013a; Bowman et al., 2014). Accordingly, in these and other previous studies dealing with spatial transferability of models, similar contexts or planning areas worthy of model transfer have been defined based on a set of exogenously specified criteria.

In this research, a latent segmentation approach to model transferability assessment is adopted so that the criteria that define segments (within which model transferability would be effective) are identified endogenously within the model estimation phase. This novel approach is applied in this paper on data from different contexts to probabilistically determine the criteria based on which transferability can be achieved. The model considered in this paper is similar to the activity generation and time-use model discussed in Sikder and Pinjari (2013a). However, rather than assessing spatial transferability via naïve transfer or transfer with constants update – as was done in their paper – this study aims to study spatial transferability in an estimation-based context where latent classification of the dataset results in achieving and endogenously identifying rather homogeneous segments comprising different regions. Model transferability is best achieved by borrowing and applying models across regions that fall into the same endogenously identified latent segment.

The remainder of this paper is organized as follows. The second section offers a description of the data set used in this study. The modeling methodology is presented in the third section. Model estimation results are presented in the fourth section, while an assessment of the latent segments and spatial transferability is furnished in the fifth section. The sixth and final section presents conclusions.

DATA

The data used in this paper is drawn from the 2009 National Household Travel Survey (NHTS), with add-on samples from 14 states and six regional and metropolitan planning agencies. The analysis is limited to considering weekday activity participation of unemployed adults (18 years or above). In order to prepare the data set for this study, extensive data filtering was performed. Records with incomplete information, missing information, weekend activity-travel records, and long distance travel (150 miles or longer) were removed from the data set. The out-of-home activities were classified into eight categories: shopping, maintenance, social/recreational, active recreation, medical, eat out, pickup/drop-off, and others. Similar activities were aggregated in terms of their dwell times. For example, if an individual performed a shopping activity for 30 minutes and another shopping activity for 50 minutes, the aggregation resulted in two shopping activities with 80 minutes of total shopping dwell time. Accordingly, the total in-home activities dwell time was inferred by subtracting the total out-of-home activities dwell time, the total travel time, and sleep time (taken to be 8.67 hours or 520 minutes[1]) from the total time of 24 hours in a day. After filtering out inconsistent records (e.g., those with dwell times and travel times adding up to more than 24 hours a day or those with combinations of dwell times and travel times that lead to negative in-home activities dwell time), and removing duplicate entries for the same individual, the final dataset included records for 28,264 individuals belonging to 39 different states. In the interest of computational time considerations, this paper focuses on weekday daily activity-travel information pertaining only to the states of California and Florida with a sample size of 10,649 individuals.

Table 1 presents the socio-economic and activity engagement characteristics of the survey data sample. The sample contains activity participation information from nine regions: Los Angeles – Riverside – Orange County, CA; Sacramento – Yolo, CA; San Diego, CA; San Francisco – Oakland – San Jose, CA; Jacksonville, FL; Miami – Fort Lauderdale, FL; Orlando, FL; Tampa – St Petersburg – Clearwater, FL; and West Palm Beach – Boca Raton, FL. The respective state samples are significantly different from one another. For example, the age distribution shows a higher percentage of young and middle aged people (18 – 54 years) in California than in Florida, a higher percentage of older individuals (55+ years) in Florida than in

[1] Based on the 2009 American Time Use Survey of the US Bureau of Labor Statistics

California. This is consistent with the notion that Florida is a popular destination for retirees and hence there is a relatively high proportion of older individuals. There is a higher percentage of people with a bachelor's degree or higher in California than in Florida. Individuals belonging to the California sample seem to be wealthier than those in the Florida sample (35.9 percent with income greater than $75,000 in California compared to 26.1 percent for Florida), although this should be interpreted in the context of the cost of living differential between the two states. Cost of living is generally higher in California than in Florida. These differences in socio-demographic characteristics between the two states may contribute to individuals residing in different areas exhibiting varying intrinsic preferences for activity participation and time-use. In view of this, it may be expected that models estimated on individual segments will perform better than a model estimated on the data set as a whole (in replicating observed activity-travel patterns in each geographical region).

The dependent variable in the modeling effort of this study is individual-level activity generation and time-use. As mentioned previously, there are eight types of out-of-home activity choices. Moreover, an individual can choose the degree to which he/she participates in the chosen activity – represented by the activity dwell time (in minutes). Table 1 shows the variability in the dependent variable characteristics across the states in the dataset. The information presented reflects the average number of activities an unemployed adult undertakes on a weekday, as well as the average duration an individual participates in a certain type of activity (by state and for the dataset as a whole). It is seen that individuals exhibit considerable similarity in their activity engagement and time use profiles, albeit with a few notable differences. For example, individuals in Florida spend more time for medical related activities (consistent with the older age profile of the survey sample), while California residents spend more time for social and other activities. Residents in California also show marginally higher levels of time use for active recreational pursuits.

MODELING METHODOLOGY

This section presents an overview of the modeling methodology adopted in this paper. The methodology includes segment-specific model formulation and assignment components that provide the ability to identify latent segments endogenously and then assign regions to different segments based on the endogenously identified criteria.

Multiple Discrete-Continuous Extreme Value Model

Single discrete choice models, such as multinomial logit (MNL) and multinomial probit (MNP), are typically utilized to model a decision making process where decision makers choose one alternative from a set of feasible alternatives. Some choice processes, however, involve the choice of multiple alternatives from the universal choice set of alternatives. An example of such a multiple-discrete choice process includes the choice of multiple vehicle types from an array of vehicle types available in the market (for example, a household may own both a car and a minivan) or the array of food choices that a household consumes. In addition to choosing multiple alternatives, an individual or household may consume each of the chosen alternatives to different degrees. Pairing multiple-discrete choice process with the continuous consumption component leads to the formulation of Kuhn-Tucker demand functions and gives rise to the multiple discrete-continuous (MDC) family of models. These models represent the decision process as a selection of one or more options from a set of alternatives, as well as the decision of the degree of consumption of the chosen alternative(s), subject to linear budget constraints. The utility function in these models is assumed to be non-linear, quasi-concave, increasing, and continuously differentiable to reflect satiation (i.e., decreasing marginal utility) as consumption increases.

The multiple discrete-continuous extreme value (MDCEV) model, proposed by Bhat (2005), accommodates multiple discreteness based on the generalized variant of the translated constant elasticity of substitution (CES) utility function with a multiplicative log-extreme value distribution for the error term. Moreover, to account for heterogeneity in the population and to produce models that better fit the available data points, population segmentation is proposed in this study. There are two methods for segmentation: exogenous and endogenous. Exogenous segmentation assumes a finite number of mutually exclusive segments, the total number of which is a function of the number of segmentation variables. An apparent setback to this approach is that the number of segments grows dramatically as the number of clustering variables increases. Endogenous segmentation, on the other hand, allows for a large number of segmentation variables to characterize each segment without having the number of segments explode. The parameters on these segmentation variables determine the propensity of belonging to each of the segments and individuals are assigned to segments in a probabilistic manner. Bhat (1997) used the endogenous segmentation approach to segment a population into a finite number of homogenous segments where the utility function is expected to be identical for all individuals probabilistically assigned to a specific segment. However, the utility function is allowed to vary across segments. The number of segments, and the variables that define the segments, are determined as part of the model estimation process. According to Bhat (1997), endogenous segmentation better fits the data as compared to exogenous segmentation, allows for higher order interaction effects, keeps the number of segments under control, and provides more intuitive results with respect to the identification of homogenous clusters of units.

In view of the above, the model used in this paper is the MDCEV model that accommodates the discrete nature of activity selection as well as the continuous nature of activity participation. To study spatial transferability, the dataset – comprising of states and regions of different socioeconomic composition – is segmented based on a number of spatial characteristics into a number of segments using latent classification. Essentially, regions belonging to the same segment, as a result of latent classification, have a unique model. In other words, parameter equality across regions of the same segment is established.

Segment-Specific Model Formulation

Assume the dataset is segmented into S homogenous segments where individuals belonging to the same segment s exhibit similar choice behavior, different than those belonging to segment s'. The model considered in this paper studies activity participation and time-use at the individual-level. All individuals participate in in-home activities and as such, in-home activities are modeled as the outside good in the model structure below – based on a generalized variant of the translated CES utility (Bhat, 2005; Bhat, 2008).

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

The first term in this expression corresponds to the utility derived from the consumption of an outside good, i.e., an alternative that is consumed by all individuals in the sample. In its absence, the expression collapses to include just the second term of Equation (1) with k ranging from 1 to K (where k is an alternative). $U_s(\boldsymbol{x})$ is the utility function associated with the consumption quantity x in segment s. It is quasi-concave, increasing, and continuously differentiable with respect to the vector $\boldsymbol{x}$ of dimension (K x 1) ($x_k \geq 0$ for all k alternatives). ψ_{ks} is the baseline marginal utility of consuming good k in segment s, i.e., the utility when there is zero consumption of good k. This utility is expressed in terms of a vector of exogenous variables $\boldsymbol{z_{ks}}$ as follows: $\psi_{ks} = \exp(\boldsymbol{\beta}'_{\boldsymbol{s}}\boldsymbol{z_{ks}} + \boldsymbol{\varepsilon_{ks}})$ where $\boldsymbol{\beta}$ is a vector of parameters reflecting the sensitivity of the baseline utility to the exogenous variables. The marginal rate of substitution between two goods i and j is the ratio of their baseline marginal utilities. Accordingly, if i and j have the same unit prices, the consumer would gain more utility consuming the alternative with the higher baseline marginal utility and is therefore, more likely to consume that good and prefer it over other goods with similar unit prices.

γ_{ks} is a parameter associated with good k in segment s and plays a dual role. On the one hand, these parameters enable corner solutions (i.e., zero consumption of a good k). On the other hand, these parameters serve as satiation parameters (reflecting preference, analogous to slopes of indifference curves). There is no translation parameter γ_{1s} associated with the outside good as it is always consumed. α_{ks} is a satiation parameter associated with good k in segment s. As more of good k is consumed, the marginal utility of additional consumption decreases. A value of one for all satiation parameters essentially implies that the consumer does not experience satiation. If there is no satiation effect and if the unit prices of all available goods are the same, the consumer is expected to invest the entirety of his or her budget in the good with the highest baseline marginal utility (i.e., the highest ψ_k value). As the value of α_k decreases from the value of unity, the satiation effect of good k increases. The inclusion of both γ_{ks} and α_{ks} in the model specification renders the estimation of Equation (1) impossible as they both reflect satiation behavior. Accordingly, $U_s(\boldsymbol{x})$ can be rewritten in two ways depending the satiation parameter that is estimated (γ_{ks} versus α_{ks}). In the case where the γ_{ks} parameters are estimated as the satiation parameters, $U_s(\boldsymbol{x})$ may be written as:

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

In the case where the α_{ks} parameters are estimated as the satiation parameters, $U_s(\boldsymbol{x})$ may be written as:

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

The first terms in equations (2) and (3) refer to the outside good, i.e., in-home activities, in the context of this paper. The MDCEV model assumes an extreme value distribution for the error term ε_{ks} and that ε_{ks} is independent of $\boldsymbol{z_{ks}}$ for all goods k. The error terms are also assumed to be independently distributed across alternatives with a scale parameter σ. However, in the absence of information on price variation across the choice alternatives, or when the price is known to be invariant across alternatives, σ can be normalized to one for convenience.

V_{ks} denotes the utility associated with alternative k in segment s and is defined based on two profiles: the γ-profile and the α-profile depicted in Equations (2) and (3) The γ-profile expression of V_{ks} is given as follows:

$$V_{ks} = \boldsymbol{\beta}'_{\boldsymbol{s}}\boldsymbol{z_{ks}} - \ln\left(\frac{x_k^*}{\gamma_{ks}} + 1\right) \quad (4)$$

Equation 4 holds for k = 2, 3, ..., K; $V_{1s} = -\ln(x_1^* + \gamma_{1s})$. The α-profile expression of V_{ks} is given as follows:

$$V_{ks} = \boldsymbol{\beta}'_s \boldsymbol{z_{ks}} + (\alpha_{ks} - 1)\ln(x_k^* + 1) \tag{5}$$

Equation 5 holds for k = 2, 3, …, K; $V_{1s} = (\alpha_{1s} - 1)\ln(x_1^*)$. Given the two profiles for the utility V_{ks}, the expression for the probability of the consumption pattern of goods k for individual q (of a total number of individuals Q) conditional on belonging to segment s is as follows:

$$P_q(x_1^*, x_2^*, x_3^*, \dots, x_M^*, 0, 0, \dots, 0)|S = \left[\prod_{i=1}^{M} f_i\right]\left[\sum_{i=1}^{M} \frac{1}{f_i}\right]\left[\frac{\prod_{i=1}^{M} e^{V_{is}}}{\left(\sum_{k=1}^{K} e^{V_{ks}}\right)^M}\right](M-1)! \tag{6}$$

where,

$$f_i = \left(\frac{1 - \alpha_i}{x_i^* + \gamma_i}\right) \tag{7}$$

M refers to the total number of consumed goods (M ≥ 1)
x_i^* refers to the consumption quantity of good i.

The individual utility maximization is subject to the budget constraint $\sum_{k=1}^{K} x_k^* = E$ where E is the total continuous quantity available to an individual (24 hours in the context of activity engagement). For convenience, the γ profile is adopted and estimated in this paper.

Segment Assignment Formulation

The latent classification aspect of this model assigns individuals (cities or regions in the context of this model) to the segments. The utility of individual q belonging to segment s is given by the following expression (Sobhani et al, 2013):

$$W_{qs}^* = \boldsymbol{\delta}'_s \boldsymbol{y_q} + \xi_{qs} \tag{8}$$

where

- $\boldsymbol{y_q}$ is a column vector (of dimension Mx1) of variables, including a constant, that influence the tendency of individual q to belong to segment s.
- $\boldsymbol{\delta_s}$ is a column vector (of dimension Mx1) of coefficients explaining the sensitivity of the utility W_{qs} to the independent variables $\boldsymbol{y_q}$.
- ξ_{qs} is an idiosyncratic random error term assumed to have an independent Type I extreme value distribution across individuals q and segments s.

Accordingly, the probability that individual q belongs to segment s is given as follows:

$$P_{qs} = \frac{\exp(\boldsymbol{\delta}'_s \boldsymbol{y_q})}{\sum_{k=1}^{S} \exp(\boldsymbol{\delta}'_k \boldsymbol{y_q})} \tag{9}$$

Building on Equations 7 and 9, the unconditional probability of the multiple-discrete continuous choice pattern is as follows:

$$P_q = \sum_{s=1}^{S} \left[(P_q(x_1^*, x_2^*, x_3^*, \dots, x_M^*, 0, 0, \dots, 0)|S * P_{qs}\right] \tag{10}$$

Consequently, the likelihood function for the entire dataset (size Q) is as follows:

$$L = \prod_{q=1}^{Q} P_q \tag{11}$$

After determining segment membership, the characteristics of each segment can be obtained by estimating the mean of the variables in each segment as follows (Bhat, 1997):

$$\overline{y_s} = \frac{\sum_q P_{qs}\boldsymbol{y_q}}{\sum_q P_{qs}} \tag{12}$$

Measures of Goodness-of-Fit

In this paper, the model is first estimated assuming the population is comprised of two segments. The number of segments is incrementally increased in a stepwise manner until further segmentation of the population no longer improves goodness-of-fit. The log-likelihood value improves as the number of segments increases, calling for the use of more effective goodness-of-fit measures for assessing the optimal number of segments in the dataset. Such measures include the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and Akaike Information Criterion corrected (AICc). The Bayesian Information Criterion (BIC) is given by the following expression (Schwarz, 1978).

$$\mathrm{BIC} = -2LL + K \ln Q \tag{15}$$

where LL is the log likelihood at convergence, K is the number of estimated parameters and Q is the number of observations in the dataset. The Akaike Information Criterion (AIC) is given by the following expression (Akaike, 1974).

$$\mathrm{AIC} = 2K - 2LL \tag{16}$$

The Akaike Information Criterion corrected ($\mathrm{AIC_c}$) is given by the following expression (Sugiura, 1978; Hurvich & Tsai, 1995).

$$\mathrm{AIC_c} = 2K - 2LL + \frac{2K(K+1)}{Q-K-1} \tag{17}$$

Several studies suggest that the BIC is superior to other assessment measures when it comes to determining the dimensionality of the segment-space (see Rust et al, 1995; Steele and Raftery, 2009). For this reason, the BIC is used in this paper as the basis for establishing the number of segments *S* into which the dataset will be divided.

LATENT SEGMENTATION RESULTS

This section presents the latent segmentation results. A base MDCEV model of activity engagement and time allocation was estimated on the entire data set. In addition, models were estimated assuming a latent segmentation with *S*=2, 3, and 4 segments. The specifications of each of these models include an array of socio-economic variables (age, gender, household size, income levels, auto ownership) and a contextual variable reflecting area type (urban/rural). The starting values for the two segments model were based on the estimation results of the base MDCEV model (estimated on the entire dataset). The starting values for the three segments model were based on the results of the two segments model and the base MDCEV model. The starting values for the four segments model were based on the results of the three segments model and the base MDCEV model. The BIC was computed for each of the models representing different segmentation schemes as per Equation (15). For the two segments model, the BIC was 326426.7. The value is found to decrease for the three segments model (326049.1), and then increase for the four segments (326053.7). Based on this finding, it may be concluded that three segments is the optimal dimension of the segment-space.

In the interest of brevity, complete model estimation results for the three-segment MDCEV model are not furnished in this paper. In general, it was found that the model estimation results are intuitive and consistent with expectations. In all segments, an array of socio-economic variables and urban area type influence activity engagement and time allocation. An examination of the signs and magnitudes of some of the variables (gender, area type, number of vehicles in the household) suggests that there is considerable heterogeneity in how individuals of different segments engage their time.

The parameters in the first segment model suggest that the activity with the highest intensity of participation is the active recreation activity. This activity entails going to the gym, exercising, and playing sports. The parameters in the second segment model show that individuals belonging to this segment engage in personal and recreational activities, namely maintenance, social, medical, and other activities, more so than individuals in other segments. The other activity category includes school-related activities, religious activities, relaxation, vacation, family obligations, attending funerals or weddings, pet care, attending meetings, and others. However, overall participation in all of the out-of-home activities in this segment remains less than the participation in in-home activities, suggesting that individuals in this segment are less out-of-home activity oriented when compared with individuals in the other two segments. The parameters in the third segment model indicate that the activity with the highest level of consumption for individuals in this segment is shopping. This includes shopping/running errands and buying goods (e.g., groceries, clothing, and hardware).

Table 2 furnishes estimation results for the latent segmentation portion of the model. This is the model that actually determines the segment into which an individual falls. Once an individual is assigned to a (latent) segment, then the appropriate MDCEV model corresponding to that segment can be used to forecast activity engagement and time use patterns for the specific individual. Residential and employment densities were used as proxies for demographic characteristics as well as area type. Urban versus rural area type is not explicitly introduced into the segmentation configuration due to a high correlation between residential and employment densities on the one hand and urban dummy variable on the other. In addition, a state-specific dummy variable was introduced to account for inter-state versus intra-state transferability. Moreover, transit service quality is represented by the presence or absence of rail in the Metropolitan Statistical Area (MSA) corresponding to the regions in the dataset. Segment 1 is treated as the base in the results furnished in Table 2. A comparison of the parameter signs and magnitudes between the second and third segments provides important qualitative information pertaining to the spatial characteristics of these two segments relative to each other. The model results yield relatively large constants for segments two and three, suggesting that these segments account for a higher share of the sample. Those residing in higher density areas are less likely to fall within segments two and three. However, those residing in high employment density locations are likely to fall within segment two. Individuals in the California data set are less likely to fall in segments two and three, suggesting that there are significant differences between the two states included in this study (Florida and California).

The quantitative characterization of the three segments in Table 2 is performed by computing the mean values of the segmentation variables within each segment as per Equation (12). Overall, it is found that the first segment accounts for about 13.5 percent of the sample, the second segment accounts for 47.7 percent of the sample, and third segment accounts for 38.8 percent of the sample (see the bottom row of Table 2). Within the context of the various characteristics, it is found that segment one is characterized by (individuals living in) areas with higher residential density, low- to medium employment density, and absence of rail service (in comparison to areas that fall into segments two and three). Consistent with this segmentation pattern, the MDCEV model estimation results show that individuals in segment one, who reside in higher residential density neighborhoods as per the segmentation model, are more likely to engage in active recreational pursuits and allocate time to such activities. Similarly, it is found that segment two is largely made up of high density residential and employment areas. The fact that the MDCEV model shows that individuals in this segment engage in a variety of activities such as personal maintenance, social/recreational activities, medical, and other can be attributed to the likelihood that such areas offer diverse and plentiful opportunities for engaging in different kinds of activities. Overall, however, those in segment two pursue out-of-home activities to a lesser degree than those in segments one and three; if they do pursue activities, then it is likely to be a variety of activities as opposed to an emphasis on just one or two activities. It is found that individuals in segment three are likely to fall into lower density areas with presumably fewer opportunities for outdoor pursuits and recreational activities. Consistent with this finding in the latent segmentation model, the MDCEV model shows that individuals in segment three are more prone to undertake shopping activities, presumably because the areas do not offer opportunities for pursuing a variety of different activities.

COMPARISON OF ENDOGENOUS AND EXOGENOUS SEGMENTATION SCHEME

This section offers a comparison of the performance of the endogenous segmentation scheme versus the traditional exogenous segmentation scheme in which segments are identified based on exogenously defined criteria. It should be noted that the adjusted log-likelihood ratio index for the three-segment MDCEV model is 0.4136 and the number of estimated parameters in the model is 325. Table 3 presents results of the comparison showing that the endogenous segmentation scheme outperforms the exogenous segmentation schemes for both one and two-way segmentations.

There are a total of 12 models in the case of a one-way exogenous segmentation scheme and a total of 57 models for all feasible combinations of two-way segmentation schemes. The best and most preferred specification for each of these models was derived by iteratively removing the insignificant parameters after every estimation run until significant and behaviorally intuitive parameters remained. The one-way segmentation model results show that area type is the most important segmentation variable with the highest adjusted likelihood ratio index among all one-way segmentation models (0.4094). For this comparison, the adjusted likelihood ratio index for all one-way segmentation models was also computed under the most favorable condition, where the index corresponds to the number of estimated parameters in the base MDCEV model (134 parameters). The resulting $\bar{\rho}^2_{fav}$ values are shown in Table 3 and indicate that, under the most favorable scenario (although unrealistic), the adjusted likelihood ratio index is still less than that of the endogenous segmentation model.

Two-way segmentation allows for higher order interaction effects and is expected to better capture the heterogeneity in preference. In each row, the two-way segmentation corresponds to the pair of variables from the left most column (one way segmentation variable) and a second variable identified in the middle column. For example, the very first row of the two-way segmentation results correspond to a segmentation based on state and residential density, the second row corresponds to a two-way segmentation based on state and employment density, and so on. The two-way segmentation model results show that the two-way segmentation model with the highest adjusted likelihood ratio index is employment density-area type (0.4084). The adjusted likelihood ratio index is then computed for the most favorable scenario where the number of parameters to be substituted in the equation resembles that of the endogenous segmentation model (325 parameters). The resulting $\bar{\rho}^2_{fav}$ values show that the residential-employment density two-way segmentation model outperforms other models (with an index of 0.4095), but still has a smaller index than that of the endogenous segmentation model (0.4136). Although these values are rather close in magnitude, the efficacy of the two-way exogenous segmentation may be suspect in view of the non-intuitive model parameter estimates obtained in that particular model estimation exercise. For example, some segments showed that individuals in the lowest income level perform more out-of-home social and recreational activities than those in the higher income levels. Another example of a counter-intuitive finding is that people older than 75 years of age engage in fewer medical activities than those belonging to the 65-74 year age group who, in turn, participate in less medical activities than those belonging to the 55-64 year age group, a result that violates a priori expectations.

It was not possible to compare the endogenous segmentation model against all possible two-way exogenous segmentation models because of small sample sizes for certain two-way segmentation schemes. This further illustrates the merits of an endogenous segmentation scheme over an exogenous segmentation scheme, the latter being limited by the available sample. By resorting to endogenous segmentation, it is possible to overcome the hurdle of small sample sizes. The endogenous segmentation scheme offered more behaviorally intuitive model parameter estimates and superior goodness-of-fit, suggesting that it outperforms other exogenous segmentation schemes adopted in prior literature.

SPATIAL TRANSFERABILITY BASED ON LATENT SEGMENTATION

The results of the modeling effort were applied to classify individuals in each of the regions into the various segments. Using the latent segmentation model, the average probability of individuals in each subregion falling into the various segments (three) is computed and tabulated in Table 4. In addition to the regions included in the estimation sample (belonging to Florida and California), the Austin-San Marcos region of Texas is included in the analysis as a region outside of the estimation sample. The segmentation model was applied to all individuals in the National Household Travel Survey (NHTS) sample of Austin-San Marcos, and the average probability of individuals falling into each of the three segments was computed. These values are shown in the last row of the table.

Using results from this table, it is possible to identify similar regions from a model spatial transferability standpoint. A pair of regions may be considered similar based on the difference in probability profile across the three latent segments. If two regions have extremely similar profiles (in terms of the proportion of individuals falling into various latent segments), then it is plausible to suggest that these two regions could spatially transfer models between one another. Using the method of least squared differences in probability profiles, it is possible to identify regions that are most similar to one another. This method was used to identify the pair of regions that are most compatible in terms of their segmentation profiles and therefore most amenable to transferring models between one another. This pairwise similarity or transferability profile is shown in Table 5. Note that the "X" in the cells of Table 5 signify region pairs that are most similar to one another.

An interesting finding is that regions from the state of Florida are more similar to other regions within the same state. The same finding holds for regions belonging to the state of California. This means that, although the segments are rather similar in their composition (with respect to the two states), intra-state transferability is still favored over inter-state transferability.

SUMMARY AND CONCLUSIONS

In an era of limited resources and ever-growing demands on disaggregate activity-travel behavior data, metropolitan planning authorities – particularly those in small and medium sized areas – are invariably interested in exploring spatial transferability of models whereby one area may transfer and apply a model estimated in a different, but similar, geographic context. Among the many (geographic) choices that are available to an area interested in borrowing (transferring) a model, which one is the best? This is the question that this paper aims to address.

The traditional approach to identifying an area with a similar profile has been to exogenously (a priori) define a limited set of criteria (say, population and employment size, level and variety of transit service), and then borrow a model from an area that has similar characteristics with respect to the chosen criteria. However, it is difficult to identify the most appropriate set of criteria a priori and the literature has utilized a variety of criteria for transferability, leaving considerable ambiguity for an agency that is seeking to transfer a model from an area with similar population activity-travel characteristics. Rather than approach the transferability paradigm through an exogenous segmentation approach, this paper proposes the utilization of an endogenous segmentation approach to help identify areas that are most similar in their profiles and thus suitable for transferring models between one another.

The model system considered in this paper is a multiple discrete-continuous extreme value (MDCEV) model of activity engagement and time allocation. Small and medium-sized communities may not have the resources to collect data and implement activity-based model systems of their own, thus calling for the potential transfer of such models from other areas. In this paper, a simultaneous equations model system approach is adopted to accommodate endogenous segmentation. The model system includes a segmentation model coupled with the segment-specific MDCEV model of activity engagement. The latent segmentation model uses a variety of explanatory factors such as area type, transit presence, residential density, and employment density to predict the segment into which an individual would fall, and then the MDCEV model can be used to predict the activity-travel pattern of an individual depending on the segment in which the individual has been placed.

In this study, a National Household Travel Survey (NHTS) sample including individuals from California and Florida is utilized to estimate the latent segmentation MDCEV model system. It is found that a three-segment model performs best in terms of goodness-of-fit and behavioral intuitiveness. The performance of the endogenous segmentation scheme is compared against the performance of alternative exogenous segmentation schemes and the endogenous segmentation scheme is found to perform consistently better than the exogenous segmentation schemes. The efficacy of the model system and approach is demonstrated through the application of the model to the regions covered by the estimation data set and one additional area not included in the estimation data set. It is found that the subregions in the data set exhibit rather similar profiles in terms of the proportion of individuals in each regional sample falling in each of the three latent segments. Using a method of least differences, it is possible to identify pairs of regions that are most closely aligned in terms of the latent segmentation profile. The results are found to be quite intuitive with intra-state transferability showing greater strength than inter-state transferability.

The paper demonstrates an approach by which areas interested in transferring models can identify similar geographic regions through an endogenous segmentation process wherein the criteria that define similarity are established within the model estimation phase. This provides a robust mechanism to identify criteria and establish similarity because the transferability assessment is being done based on those criteria that are most important in distinguishing various regions with respect to the activity-travel characteristics of interest (as identified through rigorous statistical procedures). Future research efforts should be aimed at considering alternative data sets (combining different geographical regions) and different activity-travel characteristics to explore the extent to which the criteria identified in this paper vary across data sets and activity-travel dimensions of interest.

REFERENCES

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control, 19*(6).

Arentze, T., Hofman, F., van Mourik, H., & Timmermans, H. (2002). Spatial Transferability of the Albatross Model System: Empirical Evidence from Two Case Studies. *Transportation Research Record 1805*, 1-7.

Bekhor, S., & Prato, C. G. (2009). Methodological Transferability in Route Choice Modelling. *Transportation Research Part B, 43*, 422-437.

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

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

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

Bowman, J. L., Bradley, M., Castiglione, J., & Yoder, S. L. (2014). Making Advanced Travel Forecasting Models Affordable Through Model Transferability. *Transportation Research Board 93rd Annual Meeting.* Washington, D.C.: Transportation Research Board.

Caldwell, L. C., & Demetsky, M. J. (1980). Transferability of Trip Generation Models. *Transportation Research Record 751*, 56-62.

Everett, J. D. (2009). *An Investigation of the Transferability of Trip Generation Models and the Utilization of a Spatial Context Variable.* Knoxville: University of Tennessee.

Hurvich, C. M., & Tsai, C.-L. (1995). Model Selection for Extended Quasi-Likelihood Models in Small Samples. *Biometrics*, 1077-1084.

Karasmaa, N. (2001). The Spatial Transferability of the Helsinki Metropolitan Area Mode Choice Models. *presented at the 5th workshop of the Nordic research network on modeling transport, land-use, and the environment*, (pp. 1-24).

Koppelman, F. S., & Pas, E. I. (1986). Multidimensional Choice Model Transferability. *Transportation Research Part B, 20B*(4), 321-330.

Koppelman, F. S., & Rose, J. (1983). Geographic Transfer of Travel Choice Models: Evaluations and Procedures. *Optimization and Discrete Choice in Urban Systems*, (pp. 272-309).

Koppelman, F. S., & Wilmot, C. G. (1982). Transferability Analysis of Disaggregate Choice Models. *Transportation Research Record*, *895*, pp. 18-24. Washington, D.C.

Koppelman, F. S., & Wilmot, C. G. (1986). The Effect of Omission Variables on Choice Model Transferability. *Transportation Research Part B, 20B*(3), 205-213.

McComb, L. A. (1986). Analysis of the Transferability of Sisaggregate Demand Models Among Ten Canadian Cities. *Tribune Des Transports, 3*(1), 19-32.

Mohammadian, A., & Zhang, Y. (20007). Investigating the Transferability of National Household Travel Survey Data. *Transportation Research Record 1993*, 67-79.

Nowrouzian, R., & Srinivasan, S. (2012). Empirical Analysis of Spatial Transferability of Tour-Generation Models. *Transportation Research Record 2302*, 14-22.

Reuscher, T. R., Schmoyer, R. L., & Hu, P. S. (2002). Transferability of Nationwide Personal Transportation Survey Data to Regional and Local Scales. *Transportation Research Record 1817*, 25-35.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464.

Sikder, S., & Pinjari, A. R. (2013a). Spatial Transferability of Person-Level Daily Activity Generation and Time-Use Models: An Empirical Assessment. *Transportation Research Record 2343*, 95-104.

Sikder, S., Augustin, B., Pinjari, A. R., & Eluru, N. (2013b). Spatial Transerability of Tour-based Time-of-day Choice Models: An Empirical Assessment. *Procedia - Social and Behavioral Sciences, 104*, 640-649.

Sikder, S., Pinjari, A. R., Srinivasan, S., & Nowrouzian, R. (2012). Spatial Transferability of Travel Forecasting Models: A Review and Synthesis. *4th Innovations in Travel Modeling Conference.*

Sobhani, A., Eluru, N., & Faghig-Imani, A. (2013). A Latent Segmentation Based Multiple Discrete Continuous Extreme Value Model. *Transportation Research Part B, 58*, 154-169.

Stopher, P. R., Greaves, S., & Bullock, P. (2003). Simulating Household Travel Survey Data: Application to Two Urban Areas. *Transportation Research Board 82nd Annual Meeting.* Washington, D.C.: Transportation Research Board.

Sugiura, N. (1978). Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. *Communications in Statistics, Theory and Methods, 7*, 13-26.

Wilmot, C. (1995). Evidence of Transferability of Trip Generation Models. *Journal of Transportation Engineering, 121*(5), 405-410.

Table 1. Descriptive Statistics of the Sample by State

Characteristic	California	Florida	Total
Sample Size	7,048	3,601	10,649
Gender: Male	40.7%	41.2%	40.9%
Age: 18 – 29 years	8.0%	4.3%	6.6%
Age: 30 – 54 years	22.7%	16.6%	20.6%
Age: 55 – 64 years	18.6%	18.2%	18.5%
Age: 65 – 74 years	25.5%	28.3%	26.5%
Age: ≥75 years	25.2%	32.6%	27.8%
Race: White	78.8%	87.9%	81.9%
Race: Black	3.8%	6.7%	4.8%
Race: Other	17.4%	5.4%	13.3%
Driver Status	91.3%	90.9%	91.2%
Education: High school level or lower	29.5%	37.5%	32.2%
Education: Some college level	32.1%	28.5%	30.9%
Education: Bachelor's level or higher	38.4%	34.0%	36.9%
Income: <25 K	18.5%	26.3%	21.1%
Income: 25 K – 50 K	27.7%	31.5%	29.0%
Income: 50 K – 75 K	17.9%	16.1%	17.3%
Income: ≥75 K	35.9%	26.1%	32.6%
Average Household Size	2.5	2.2	2.4
Average Number of Drivers	1.9	1.8	1.9
Average Number of Activities	3.0	3.0	3.0
Average Activity Duration (min)[a]			
Home	702.5	705.8	703.6
Shop	60.3	61.1	60.6
Maintenance	31.3	31.8	31.5
Social	161.9	156.7	160.1
Active	83.7	79.8	82.4
Medical	79.2	87.6	82.0
Eat-out	63/4	65.2	64.0
Pick-up	44.7	43.6	44.3
Other	148.9	121.1	139.5

[a]average durations are computed only on the portion of the sample that participated in each of the activities. Table

Table 2. The Latent Segmentation Model and Characterization of Three Segments

Segmentation Variable		Segment 1 (base)	Segment 2	Segment 3	Dataset
Constants		-	1.6016 (14.56)	1.4904 (13.27)	-
Residential Density (Housing units per sq mi)	< 500 (base)	-	-	-	-
	500 – 1,999	-	-0.2127 (-2.18)	-0.2089 (-2.12)	-
	≥ 2,000	-	-0.1324 (-1.33)	-0.1710 (-1.72)	-
Employment Density (Workers per sq mi)	< 500 (base)	-	-	-	-
	500 – 1,999	-	-	-	-
	≥ 2,000	-	0.1487 (2.70)	-	-
State	California	-	-0.2391 (-3.38)	-0.2709 (-3.74)	-
	Florida (base)	-	-	-	-
Transit Service Quality	Rail (base)	-	-	-	-
	No Rail	-	-0.1450 (-2.18)	-0.1805 (-2.66)	-
Quantitative Characterization of the Three Segments					
Residential Density (Housing units per sq mi)	< 500	14.07%	15.53%	16.31%	15.64%
	500 – 1,999	42.14%	38.99%	40.27%	39.91%
	≥ 2,000	43.79%	45.48%	43.42%	44.45%
Employment Density (Workers per sq mi)	< 500	32.50%	31.84%	33.48%	32.57%
	500 – 1,999	40.89%	38.34%	39.88%	39.28%
	≥ 2,000	26.61%	29.82%	26.64%	28.15%
State	California	70.00%	65.94%	65.16%	66.18%
	Florida	30.00%	34.06%	34.84%	33.82%
Transit Service Quality	Rail	47.46%	50.47%	50.27%	50.00%
	No Rail	52.54%	49.53%	49.73%	50.00%
Area Type	Urban	93.10%	92.60%	92.26%	92.53%
	Rural	6.90%	7.40%	7.74%	7.47%
Share		0.1351	0.4771	0.3878	1.0000

Table 3. Comparison of the Endogenous Segmentation Model with One-Way and Two-Way Exogenous Segmentation Models

Segmentation Variable	Segs	Params	LL at converg	$\bar{\rho}^2$[2]	$\bar{\rho}^2_{fav}$[3]	Two-way Segmentation with…	Segs	Params	LL at converg	$\bar{\rho}^2$	$\bar{\rho}^2_{fav}$[4]
State	2	237	-163160.4833	0.4079	0.4083	Residential Density	6	563	-162805.1273	0.4080	0.4089
						Employment Density	6	597	-147394.6772	0.4079	0.4090
						Transit Service Quality	4	401	-162943.5897	0.4081	0.4084
						Area Type	4	372	-162993.2865	0.4080	0.4082
Residential Density	3	338	-163044.4900	0.4080	0.4087	Employment Density[5]	9	705	-160191.1947	0.4081	0.4095
						Transit Service Quality	6	580	-162790.0652	0.4080	0.4089
						Area Type	6	409	-160412.7488	0.4082	0.4085
Employment Density	3	350	-163032.0739	0.4080	0.4088	Transit Service Quality	6	601	-162768.6610	0.4080	0.4090
						Area Type	6	439	-161115.1576	0.4084	0.4088
Transit Service Quality	2	238	-163159.836	0.4079	0.4083	Area Type	4	304	-159485.0556	0.4082	0.4081
Area Type	2	213	-163165.1216	0.4094	0.4096	-	-	-	-	-	-

Note: The adjusted log likelihood ratio index for the 3 segments model is 0.4136 and the number of estimated parameters in this model is 325.

[2] The adjusted log likelihood ratio index is computed as $\bar{\rho}^2 = 1 - \frac{\text{LL at convergence} - \text{k}}{\text{LL at zero}}$ where k is the number of estimated parameters.

[3] The favorable adjusted log likelihood ratio index for the one-way segmentation models is computed by replacing *k* with the number of estimated parameters in the unsegmented model.

[4] The favorable adjusted log likelihood ratio index for the two-way segmentation models is computed by replacing *k* with the number of estimated parameters in the three-segments model.

[5] The two-way segmentation models between the lowest level of residential density (< 500 housing units per square mile) and the middle and highest level of employment density (500 – 1,999 and ≥ 2,000 workers per square mile respectively), the two-segmentation models between employment density and rural area type, and the two-way segmentation model between the rail transit service quality and the rural area type were not estimable due to small sample size.

Table 4. Propensity of Belonging to Each Segment by Region

Region (q)	Sample Size	P_{q1}	P_{q2}	P_{q3}
Jacksonville, FL	494	0.1256	0.4718	0.4026
Los Angeles – Riverside – Orange County, CA	2635	0.1343	0.4795	0.3862
Miami – Fort Lauderdale, FL	1183	0.1078	0.4916	0.4006
Orlando, FL	464	0.1251	0.4733	0.4016
Sacramento – Yolo, CA	471	0.1546	0.4689	0.3765
San Diego, CA	2437	0.1556	0.4706	0.3738
San Francisco – Oakland – San Jose, CA	1505	0.1338	0.4777	0.3885
Tampa – St Petersburg – Clearwater, FL	1010	0.1254	0.4765	0.3981
West Palm Beach – Boca Raton, FL	450	0.1274	0.4776	0.3950
Austin – San Marcos, TX	**568**	**0.1204**	**0.4766**	**0.4030**

Table 5. Matrix of Similarity for Regions Within and External to the Dataset

Borrow From	**Jacks**	**LA – Rivers - OC**	**Miami – Fort Laud**	**Orlan**	**Sac – Yolo**	**San Diego**	**San Fran – Oak – San Jose**	**Tampa – St Pete – Clearw**	**West Palm Beach – Boca Raton**
Jacksonville									X
LA – Riverside - OC							X		
Miami – For Laud								X	
Orlando								X	
Sacramento – Yolo						X			
San Diego					X				
San Fran – Oakland – San Jose		X							
Tampa – St Pete – Clearwater				X					
West Palm Beach – Boca Raton	X								
Austin – San Marcos								X	

X denotes the closest similarity between a region pair and the best candidate for model transferability.