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: 404-385-3754; Fax: 404-894-2278
Email: ram.pendyala@ce.gatech.edu

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

Revised March 15, 2015

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. 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 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. The end products are a set of optimal criteria for clustering regions as well as a fully transferred model, segmented to account for heterogeneity in the population. The methodology is demonstrated and its efficacy established through a case study in this paper that utilizes the National Household Travel Survey (NHTS) dataset for information on weekday activities of non-workers within nine regions in the states of California and Florida. The estimated model is then applied to a context withheld from the original estimation to assess its performance. It is found that the methodology offers a robust mechanism for identifying latent segments and establishing criteria for transferring models between areas.

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

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 approximate locally observed behavior in the application context. While it is generally considered good practice to develop models based on locally collected data, 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 (*1*). When such model transfer is considered, it is important to ensure that the transferred model offers useful and valid information in the application context (*2*).

Traditionally, in the absence of any local data, the transfer method is based on identifying another metropolitan area that is similar to the local context (*3-13*). The "similarity" between the local region and the donor region (from which the model is borrowed or transferred) is based on factors such as transit service quality (*14-16*), metropolitan area size (*14*), metropolitan region density and type (*14, 15, 18, 19*), whether the donor region is in the same state as the local region (*10, 13*), and demographic characteristics (*16, 17*). The problem with this approach to model transfer is three-fold: (1) It specifies a priori the parameter(s) that define the similarity between the local region and the donor region; (2) It assumes that a single uniform set of parameters measuring similarity are at work regardless of the type of model component being transferred; and (3) The transfer is based on centralized measures of tendency (averages, medians) between the local region and the donor region.

The first problem refers to the fact that the parameters of similarity are exogenously identified. However, it is likely that similarity between regions is a multi-dimensional measure. One way to accommodate this multi-dimensional similarity within this exogenous approach is to partition regions along all potentially relevant dimensions. However, a practical problem with this "full-dimensional" exogenous transfer scheme is that there may not be a unique region that lies at the intersection of all of the dimensions as the local region. To overcome this limitation, it is typical to consider only one or two dimensions that are a priori designated as the most important measures of closeness (similarity). The disadvantage is that closeness on a whole set of potentially important dimensions is discarded and lost. In addition, an intrinsic problem with all exogenous transfer approaches is that the threshold values of the continuous variables (for example, residential or employment density) have to be established in a rather ad-hoc fashion.

The second problem is that the exogenous approach uses the same set of similarity dimensions regardless of the type of model being transferred. In reality, it is possible that residential density is a better measure of similarity when transferring a model associated with activity time use behavior, while the availability of specific forms of transit as in the local region may be the key similarity measure for transferring a mode choice model. What is needed is a method to extract information regarding similarity in a way that is customized to the model component being transferred.

The third problem is that there are likely to be different spatial pockets within metropolitan areas that are quite different from one another on the similarity measures used in the exogenous schemes. However, the exogenous schemes use a single central measure to characterize entire metropolitan regions (such as a mean residential density measure), and use that central tendency to determine the region that is most similar to the local region. However, the local region may have pockets that are highly dense that reveal individual and household-

level activity-travel behavior patterns similar to dense pockets in other regions, while also having pockets of low density in which the activity-travel behavior patterns are similar to low density pockets in other regions. What is needed is a method of model transferability that accommodates the heterogeneity in locational characteristics and associated behaviors within the local region.

This paper presents a new latent-segmentation based endogenous approach to model transferability that overcomes the three key problems described above. In this approach, data is utilized from all of the regions that have information on the activity-travel dimensions of interest (and appropriate exogenous variables), rather than utilizing data from a single region that is chosen a priori as the single most similar region to the local context in question. In this latent segmentation based endogenous approach, there is no need to limit the dimensions of similarity to one or two, because the concept of similarity is simultaneously based on multiple dimensions. In particular, a limited number of latent segments is derived, specific to each kind of model component being transferred, by characterizing each latent segment by the entire set of potentially relevant similarity variables. The number of latent segments that is appropriate for a specific activity-travel dimension of interest is determined statistically by successively adding an additional segment until a point is reached where an additional segment does not result in a significant improvement in fit. Individuals, based on their location characteristics as captured in the potentially relevant similarity variable measures, are assigned to segments in a probabilistic fashion. That is, each latent segment refers to an optimal combination of location characteristics that make individuals within that segment behave similarly on the activity-travel dimension of interest. The endogenous approach jointly determines the number of segments, the assignment of individuals to segments, and segment-specific choice model parameters. Since this approach identifies segments without requiring a multi-way partitioning based on all potentially relevant similarity variables as in the full-dimensional exogenous transfer method, it allows the use of all similarity variables in practice. Because the similarity-based latent segmentation scheme is estimated jointly with the main activity-travel dimension model of interest, it is immediately customized to the local population context. Finally, by using data from a host of different regions (for which data is available), it is possible to capture the heterogeneity in locational characteristics and the impact of such heterogeneity on activity-travel behavior dimensions of interest. This allows the recognition of heterogeneity that exists in different spatial pockets in the local region.

The activity-travel model considered in this paper is similar to the activity generation and time-use model discussed in Sikder and Pinjari (*10*). However, rather than assessing spatial transferability via naïve transfer or transfer with constants update – as was done in their paper – this study establishes spatial transferability in an estimation-based context such that latent classification of the dataset results in endogenously identifying appropriate segments that are homogeneous with respect to the activity-travel dimension of interest.

The remainder of this paper is organized as follows. The second section offers a description of the dataset 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 NHTS. The analysis considers weekday activity participation of unemployed adults (18 years or above). In order to prepare the dataset 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 dataset. 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. 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 520 minutes according to the 2009 American Time Use Survey) from 24 hours in a day. After filtering out inconsistent records (those with dwell times and travel times adding up to more than 24 hours a day and 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, and a higher percentage of older individuals (55+ years) in Florida than in California. This is consistent with the notion that Florida is a popular destination for retirees. Individuals belonging to the California sample seem to be wealthier than those in the Florida sample, although this should be interpreted in the context of the cost of living differential between the two states. 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.

The dependent variable in this study is individual-level activity generation and time-use. As mentioned previously, there are eight types of out-of-home activities. 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 in the context of modeling an activity-travel dimension of interest.

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. The multiple discrete-continuous extreme value (MDCEV) model, proposed by Bhat (*20*), 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.

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. As mentioned earlier, 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 (*21*) 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. It is found that 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 (*21*).

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. The model jointly determines the mix of activities (multiple discrete choices) that an individual undertakes in a day together with the amount of time (continuous choice dimension) that is dedicated to each activity type. To study spatial transferability, the dataset – comprising of states and regions of different socioeconomic composition – is segmented using a number of explanatory variables that facilitate latent classification.

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 (*20*, *22*).

$$U_q(\boldsymbol{x}) \mid q \in \text{location segment } s = \exp(\varepsilon_{q1s}) \ln\{x_{q1} + \gamma_{1s}\} + \sum_{k=2}^{K} \gamma_{ks} \exp(\boldsymbol{\beta}_s' \boldsymbol{z}_{qk} + \varepsilon_{qks}) \ln\left\{\frac{x_{qk}}{\gamma_{ks}} + 1\right\}, \quad (1)$$

where,

$U_q(\boldsymbol{x}) \mid q \in \text{location segment } s$ is the utility accrued by individual q, given s/he belongs to location segment s,

k is an index for alternatives,

ε_{qks} terms are random error terms,

γ_{ks} terms are satiation parameters (discussed further below),

$\boldsymbol{\beta}_s$ is a parameter vector specific to location segment s,

$\boldsymbol{z}_{qk}$ is a set of exogenous covariates specific to individual q and alternative k, and

x_{qk} is the consumption amount of good k.

The first term in Equation 1 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 (i.e., in-home activities, in the context of this paper). 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). $U_q(\boldsymbol{x}) \mid q \in \text{location segment } s$ is quasi-concave, increasing, and continuously differentiable with respect to the vector $\boldsymbol{x}$ of dimension ($K \times 1$) ($x_k \geq 0$ for all k alternatives). 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.

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.

The expression for the probability of the consumption pattern for individual q conditional on belonging to segment s is as follows:

$$P_q(x_1^*, x_2^*, x_3^*, ..., x_M^*, 0, 0, ..., 0) \mid q \in \text{location segment } s$$

$$= \left[\prod_{i=1}^{M} f_i\right]\left[\sum_{i=1}^{M} \frac{1}{f_i}\right]\left[\frac{\prod_{i=1}^{M} e^{V_{qis}}}{\left(\sum_{k=1}^{K} e^{V_{qks}}\right)^M}\right](M-1)! \quad (2)$$

where,

$$V_{q1s} = -\ln(x_{qi}^{*} + \gamma_{is}),$$

$$V_{qks} = \boldsymbol{\beta}_{s}' \boldsymbol{z}_{\boldsymbol{qs}} - \ln\left(\frac{x_{qk}^{*}}{\gamma_{ks}} + 1\right),$$

$$f_i = \left(\frac{1-\alpha_i}{x_i^{*} + \gamma_i}\right), \tag{3}$$

M refers to the total number of consumed goods (M ≥ 1), and
x_{qk}^{*} refers to the observed consumption quantity of good k by individual q.

The individual utility maximization is subject to the budget constraint $\sum_{k=1}^{K} x_{qk}^{*} = E_q$, where E_q is the total continuous quantity available to individual q.

Segment Assignment Formulation

The latent classification aspect of this model assigns individuals to segments. The utility of individual q belonging to segment s is given by the following expression (*23*):

$$W_{qs}^{*} = \boldsymbol{\delta}_{s}' \boldsymbol{y}_{\boldsymbol{q}} + \xi_{qs}, \tag{4}$$

where,

W_{qs}^{*} is the latent propensity of individual q to belong to location segment s,
$\boldsymbol{y}_{\boldsymbol{q}}$ is a vector of individual specific exogenous covariates (including characteristics of location of the individual),
$\boldsymbol{\delta}_{\boldsymbol{s}}$ is a vector of coefficients, and
ξ_{qs} is a random error term (Type 1 extreme value).

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

$$P_{qs} = \frac{\exp(\boldsymbol{\delta}_{s}' \boldsymbol{y}_{q})}{\sum_{s'=1}^{S} (\boldsymbol{\delta}_{s'}' \boldsymbol{y}_{q})}. \tag{5}$$

Building on Equations 2 and 5, the unconditional probability of the multiple-discrete continuous choice pattern is as follows:

$$P_q = \sum_{s=1}^{S} \left[\left\{ P_q(x_1^{*}, x_2^{*}, x_3^{*}, ..., x_M^{*}, 0, 0, ..., 0) \mid q \in \text{location segment } s \right\} \times P_{qs} \right]. \tag{6}$$

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

$$L = \prod_{q=1}^{Q} P_q. \tag{7}$$

After determining segment membership, the characteristics of each segment can be derived by estimating the mean of the variables in each segment as follows (*21*):

$$\bar{y}_s = \frac{\Sigma_q P_{qs} \boldsymbol{y_q}}{\Sigma_q P_{qs}}. \quad (8)$$

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 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 (*24*).

$$\text{BIC} = -2LL + K \ln Q, \quad (9)$$

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 (*25*).

$$\text{AIC} = 2K - 2LL. \quad (10)$$

The Akaike Information Criterion corrected (AIC_c) is given by the following expression (*26*, *27*).

$$\text{AIC}_C = 2K - 2LL + \frac{2K(K+1)}{Q-K-1}. \quad (11)$$

Several studies suggest that the BIC is superior to other assessment measures when it comes to determining the dimensionality of the segment-space (see *28*, *29*). 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 was estimated on the entire dataset. 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 9. 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 model

(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, the paper does not include all tabulations of the MDCEV model estimation results. In general, it was found that the model estimation results are intuitive and consistent with expectations. 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 use 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/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.

Table 2 furnishes estimation results for just the latent segmentation portion of the model, which determines the segment into which an individual falls. Once an individual is assigned to a segment, then the MDCEV model corresponding to that segment can be used to forecast activity engagement and time use patterns for that individual. Residential and employment densities were used as proxies for area type due to a high correlation between residential and employment densities on the one hand and the urban dummy variable on the other. In addition, a state-specific dummy variable was introduced to account for differences among individuals across states. 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 dataset are less likely to fall in segments two and three, suggesting that there are significant differences between the California and Florida samples.

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 8. 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 the third segment accounts for 38.8 percent of the sample. 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 individuals 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 activities. Similarly, it is found that segment two is largely made up of individuals in high density residential and employment areas. The fact that the MDCEV model shows that individuals in this segment engage in a variety of activities can be attributed to the likelihood that such areas offer diverse and plentiful opportunities for engaging in different kinds of activities. Also, 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, 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 joint three-segment latent class 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 exogenous segmentation schemes for both one and two-way segmentations.

Traditionally, clusters have been defined by predetermined criteria in order to transfer models between areas. Twelve possible clusters emerge if segments are to be exogenously defined based on one clustering criterion. The expansion of the segmentation dimensionality to two explodes the number of possible clusters into 57, accounting for all feasible combinations of two-way segmentation. The preferred specification for each of these exogenously segmented models was derived iteratively until significant and behaviorally intuitive parameters remained. The adjusted likelihood ratio indexes for the one-way segmentation models were calculated, all of which were lower than that of the three-segment MDCEV model. Moreover, the adjusted likelihood ratio index for all one-way segmentation models was 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 preference heterogeneity. 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 corresponds to a segmentation based on state and residential density, the second row corresponds to a segmentation based on state and employment density, and so on. The adjusted likelihood ratio indexes for all two-way segmentation models were calculated using the number of parameters estimated in the corresponding models as well as using the number of parameters estimated in the endogenous segmentation model (most favorable condition). The resulting $\bar{\rho}^2$ and $\bar{\rho}^2_{fav}$ are less than the adjusted likelihood ratio index of the endogenous segmentation model. In addition, 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 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 is unlikely that only one characteristic of a geographic region deems it similar to another and justifies model transferability between them. Resorting to a higher order segmentation scheme presents issues with the number of segments under consideration as well as the sample size within each segment. In fact, 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. Also, 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

In practice, a model from a specific area is often transferred to the local context. As mentioned earlier, the model tends to be transferred from a metropolitan area that has similar traits along one or two key dimensions (that are defined a priori exogenously). In this section, the metropolitan area of Austin – that was not included in the model estimation phase – is considered as a local context. National household travel survey data are available for Austin, making it possible to estimate a MDCEV model of activity engagement and time use native to Austin. In addition, various models are "transferred" to Austin (under the hypothetical scenario that local data are not available for Austin). Considering the spatial and transit characteristics of Austin, models were estimated and transferred from the metropolitan area samples of Orlando, West Palm Beach, and Sacramento. In other words, the Orlando (or West Palm Beach or Sacramento) sample was treated as an exogenously defined segment, a MDCEV model was estimated on that segment (sample), and the parameters were transferred to the Austin sample (a constrained parameter estimation was performed on the Austin sample, where the values of all parameters are set to the Orlando/West Palm Beach/Sacramento parameter values). The latent segmentation based MDCEV model was also transferred to Austin, where individuals in the Austin sample are assigned to a latent segment and a segment-specific MDCEV model is applied accordingly.

The results of the comparison are shown in Table 4. The goodness-of-fit for the native Austin estimation is 0.4133. The latent segmentation model (which was estimated on California and Florida samples), when transferred to the Austin sample, offers a goodness of fit of 0.4125, which is very close to the goodness-of-fit of the native Austin model. All other transferred models present goodness-of-fit statistics that are less in magnitude than that offered by the latent segmentation model. In addition, it is striking to note that the log-likelihood at convergence for the transferred Orlando, West Palm Beach, and Sacramento models are all worse than the corresponding log-likelihood values with constants. In other words, transferring a number of parameters (associated with exogenous variables that explain activity-travel engagement and time use) from the respective metro areas to Austin results in a degradation of the log-likelihood; however, this is not the case for the latent segmentation model where the log-likelihood at convergence is better than that corresponding to the model with only constants. It should be noted, however, that this comparison is qualitative in nature; as the number of parameters between the latent segmentation model and the other models is considerably different, more rigorous comparisons of goodness-of-fit measures (that adjust for number of parameters) need to

be undertaken. Further validation of the efficacy of the latent segmentation approach would constitute fruitful directions for future research.

SUMMARY AND CONCLUSIONS

In an era of limited resources and ever-growing demands on disaggregate activity-travel behavior data, metropolitan planning authorities are invariably interested in exploring spatial transferability of models in order to achieve time and cost savings. The traditional approach to identifying an area with a similar profile has been to exogenously define a limited set of criteria 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 which provides the ability to classify individuals into relatively homogenous segments and then apply segment-specific activity-travel model components to simulate travel demand.

In this paper, a simultaneous equations model system approach is adopted to accommodate endogenous segmentation. The model system includes a segmentation model coupled with a segment-specific econometric model of activity participation and time use. The latent segmentation model uses a variety of explanatory factors such as area type, rail presence, residential density, and employment density to predict the segment to which an individual belongs, and then the econometric multiple discrete continuous extreme value (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 found to be consistently better than alternative exogenous segmentation schemes. The efficacy of the approach is demonstrated through the transfer of the model onto a region that was not part of the estimation dataset, the Austin-San Marcos metropolitan area. The comparison of the predicted time allocations of a locally estimated MDCEV model with that of the transferred latent segmentation model reveals similar predictive powers in replicating observed activity consumptions in the Austin-San Marcos dataset, suggesting that the segmented model offers results comparable to those obtained from estimating models on local data.

The paper demonstrates an approach for model transferability through an endogenous segmentation process wherein the criteria that define segments are established within the model estimation phase. This provides a robust mechanism to identify criteria and establish segment-specific model parameters with respect to the activity-travel characteristics of interest. Future research efforts should be aimed at considering alternative datasets (combining different geographical regions) and different activity-travel characteristics to further validate the approach.

ACKNOWLEDGEMENTS

This research was partially supported by the U.S. Department of Transportation through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center.

REFERENCES

1. Sikder, S., A. R. Pinjari, S. Srinivasan, and R. Nowrouzian. Spatial Transferability of Travel Forecasting Models: A Review and Synthesis. *4th Innovations in Travel Modeling Conference*, 2012.
2. Koppelman, F. S., and C. G. Wilmot. Transferability Analysis of Disaggregate Choice Models. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 895*, Transportation Research Board of the National Academies, Washington, D.C., 1982, pp. 18–24.
3. Koppelman, F. S., and J. Rose. Geographic Transfer of Travel Choice Models: Evaluations and Procedures. In *Optimization and Discrete Choice in Urban Systems* (B. G. Hutchinson, P. Nijkamp, and M. Batty, eds), Springer-Verlag, Berlin Heidelberg, 1985, pp. 272–309.
4. Koppelman, F. S., and C. G. Wilmot. The Effect of Omission Variables on Choice Model Transferability. *Transportation Research Part B*, Vol. 20, No. 3, 1986, pp. 205–213.
5. Koppelman, F. S., and E. I. Pas. Multidimensional Choice Model Transferability. *Transportation Research Part B*, Vol. 20, No. 4, 1986, pp. 321–330.
6. Wilmot, C. Evidence of Transferability of Trip Generation Models. *Journal of Transportation Engineering*, Vol. 121, No.5, 1995, pp. 405-410.
7. Arentze, T., F. Hofman, H. van Mourik, and H. Timmermans. Spatial Transferability of the Albatross Model System: Empirical Evidence from Two Case Studies. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1805*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 1–7.
8. Bekhor, S., and C. G. Prato. Methodological Transferability in Route Choice Modelling. *Transportation Research Part B*, Vol. 43, 2009, pp. 422-437.
9. Nowrouzian, R., and S. Srinivasan. Empirical Analysis of Spatial Transferability of Tour-Generation Models. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2302*, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 14–22.
10. Sikder, S., and A. R. Pinjari. Spatial Transferability of Person-Level Daily Activity Generation and Time-Use Models: An Empirical Assessment. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2343*, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 95–104.
11. Sikder, S., B. Augustin, A. R. Pinjari, and N. Eluru. Spatial Transferability of Tour-based Time-of-day Choice Models: An Empirical Assessment. *Procedia - Social and Behavioral Sciences*, Vol. 104, 2013, pp. 640–649.
12. Karasmaa, N. 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, 2001, pp. 1–24.
13. Bowman, J. L., M. Bradley, J. Castiglione, and S. L. Yoder. Making Advanced Travel Forecasting Models Affordable Through Model Transferability. Presented at 93rd Annual Meeting of the Transportation Research Board. Washington, D.C., 2014.
14. McComb, L. A. Analysis of the Transferability of Disaggregate Demand Models Among Ten Canadian Cities. *Tribune Des Transports*, Vol. 3, No. 1, 1986, pp. 19–32.
15. Stopher, P. R., S. Greaves, and P. Bullock. Simulating Household Travel Survey Data: Application to Two Urban Areas. Presented at 82nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2003.

16. Mohammadian, A., and Y. Zhang. Investigating the Transferability of National Household Travel Survey Data. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1993*, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 67–79.
17. Caldwell, L. C., and M. J. Demetsky. Transferability of Trip Generation Models. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 751*, Transportation Research Board of the National Academies, Washington, D.C., 1980, pp. 56–62.
18. Reuscher, T. R., R. L. Schmoyer, and P. S. Hu. Transferability of Nationwide Personal Transportation Survey Data to Regional and Local Scales. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1817*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 25–35.
19. Everett, J. D. An Investigation of the Transferability of Trip Generation Models and the Utilization of a Spatial Context Variable. University of Tennessee, Knoxville, 2009.
20. Bhat, C. R. A Multiple Discrete-Continuous Extreme Value Model: Formulation and Application to Discretionary Time-Use Decisions. *Transportation Research Part B*, Vol. 39, No. 8, 2005, pp. 679–707.
21. Bhat, C. R. An Endogenous Segmentation Mode Choice Model with an Application to Intercity Travel. *Transportation Science*, Vol. 31, No. 1, 1997, pp. 34–48.
22. Bhat, C. R. The Multiple Discrete-Continuous Extreme Value (MDCEV) Model: Role of Utility Function Parameters, Identification Considerations, and Model Extensions. *Transportation Research Part B*, Vol. 42, No. 3, 2008, pp. 274–303.
23. Sobhani, A., N. Eluru, and A. Faghig-Imani. A Latent Segmentation Based Multiple Discrete Continuous Extreme Value Model. *Transportation Research Part B*, Vol. 58, 2013, pp. 154–169.
24. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461–464.
25. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, Vol. 19, 1974, pp. 716–723
26. Sugiura, N. Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. *Communications in Statistics, Theory and Methods*, Vol. 7, 1978, pp. 13–26.
27. Hurvich, C. M., and C.-L. Tsai. Model Selection for Extended Quasi-Likelihood Models in Small Samples. *Biometrics*, Vol. 51, No. 3, 1995, pp. 1077–1084.
28. Rust, R., D. Simester, R. Brodie, and V. Nilikant. Model Selection Criteria: An Investigation of Relative Accuracy, Posterior Probabilities, and Combinations of Criteria. *Management Science*, Vol. 41, No. 2, 1995, pp. 322–333.
29. Steele, R., and A. Raftery. Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models. University of Washington, Seattle, 2009.
30. Koppelman, F. S., and C. Bhat. A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models. U.S. Department of Transportation Federal Transit Administration, 2006.

LIST OF TABLES

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 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	$\bar{\rho}^2$ [a]	$\bar{\rho}^2_{fav}$ [b]	Two-way Segmentation with…	Segs	Params	LL at converg	$\bar{\rho}^2$	$\bar{\rho}^2_{fav}$ [c]
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[d]	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.

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

[b] 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.

[c] 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.

[d] 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 Comparison of Goodness-of-Fit for Different Model Transfer Schemes

	MDCEV estimated on Austin	LC model transferred to Austin	Orlando model transferred to Austin	West Palm Beach model transferred to Austin	Sacramento model transferred to Austin
Number of Parameters	80	325	82	75	76
Log-likelihood at zero	-14752.6	-14752.6	-14752.6	-14752.6	-14752.6
Log-likelihood at constant	-8818.4	-8736.8	-8842.1	-8847.1	-8831.5
Log-likelihood at convergence	-8655.1	-8666.8	-8878.5	-8995.6	-8846.8
Rho-Squared w.r.t. Zero	0.4133	0.4125	0.3982	0.3902	0.4003