Noname manuscript No.
(will be inserted by the editor)

Assessing Preference Heterogeneity for Mobility-on-Demand Transit Service in Low-Income Communities: A Latent Segmentation Based Decision Tree Method

Xilei Zhao · Xinyi Wang · Xiang Yan · Zhuoxuan Cao

Received: date / Accepted: date

Abstract The future of public transit service is often envisioned as Mobility-on-Demand (MOD), i.e., a system that integrates fixed routes and shared on-demand shuttles. The MOD transit system has the potential to provide better transit service with higher efficiency and coverage. However, little research has focused on understanding traveler preferences for MOD transit and preference heterogeneity, especially among the disadvantaged population. This study addresses this gap by proposing a two-step method, called latent segmentation based decision tree (LSDT). This method first uses a latent class cluster analysis (LCCA) that extracts traveler profiles who have similar usage patterns for shared modes. Then, decision trees (DT) are adopted to reveal the associations between various factors with preferences for MOD transit across different clusters. We collected stated-preference data among two

Xilei Zhao, Ph.D., Corresponding Author
Department of Civil and Coastal Engineering
University of Florida
1949 Stadium Rd, Gainesville, FL 32611
Tel.: +1 (352) 294-7159
E-mail: xilei.zhao@essie.ufl.edu
ORCID: https://orcid.org/0000-0002-7903-4806

Xinyi Wang
School of Civil and Environmental Engineering
Georgia Institute of Technology
E-mail: xinyi.wang@gatech.edu
ORCID: https://orcid.org/0000-0002-3564-9147

Xiang Yan, Ph.D.
Department of Civil and Coastal Engineering
University of Florida
E-mail: xiangyan@ufl.edu
ORCID: https://orcid.org/0000-0002-8619-0065

Zhuoxuan Cao
Department of Civil and Coastal Engineering
University of Florida
E-mail: zhuoxuancao@ufl.edu
ORCID: https://orcid.org/0000-0002-3247-1291

low-resource communities, i.e., Detroit and Ypsilanti, Michigan. The LCCA model divides the entire sample into three clusters, i.e., shared-mode users, shared-mode non-users, and transit-only users. We find that job accessibility by transit is the most important variable for all the cluster-specific DT to model the MOD transit preference, and it negatively associated with the MOD transit preference. For transit-only users, gender and car ownership are the second-important variables, but neither of them appears in the DT for the other two clusters. In particular, female transit-only users have lower preference for MOD transit, possibly due to safety concerns. The LSDT method can generate richer insights than a single DT fitted to the overall sample by better accounting for heterogeneity. The findings gained from this approach can inform better-targeted strategies to plan for MOD transit services.

Keywords Mobility-on-Demand Transit · Heterogeneity · Latent class cluster analysis · Decision trees · Low-income

1 Introduction

In recent years, many transit observers have envisioned the future of transit to be a Mobility-on-Demand (MOD) transit system that integrates fixed-route services with on-demand ridesharing (Maheo et al., 2019; Shen et al., 2018; Yan et al., 2019b). The MOD transit system may enhance transit operations by solving the first-/last-mile problems, filling in the gaps in the existing services, enhancing accessibility for under-served communities, increasing transit ridership, and cutting operational costs.

To better plan and implement the MOD transit system, it is essential to study traveler preferences for MOD transit and preference heterogeneity, especially among the disadvantaged populations (who are often low-income, less-educated, carless, elderly, etc.). These disadvantaged individuals are usually more transit-dependent, but are more likely to have low technological capability and lack access to smartphones or data plans (Pew Research Center, 2018). Therefore, it is imperative to study the needs of disadvantaged travelers to better inform policies and strategies. However, few published studies have focused on this topic. To fill this research gap, in this study we address the following research questions (RQs):

- **RQ 1:** *What travel profiles can we extract from individuals living in low-income communities based on their current use of transit and ridehailing?*
- **RQ 2:** *What factors (e.g., demographic and socioeconomic characteristics, and built-environment variables) are associated with traveler preferences for MOD transit and how do these associations differ across traveler profiles?*

To answer these questions, we adopt a latent segmentation based decision tree (LSDT) method. The LSDT method includes two steps, namely, (1) applying latent class cluster analysis (LCCA) to segment the market by using travelers' current bus and ridehailing usage as the indicators, and (2) probabilistically assigning travelers to each cluster (i.e., a traveler can be 20% in Cluster 1 and 80% in Cluster 2, if there are two clusters in total), and fitting different decision tree (DT) models to different clusters. Each step answers a **RQ** discussed above.

Two-step methods like LSDT have been applied in the field of transportation to better account for heterogeneity. For example, a similar approach[1], i.e., LCCA plus DT, has been used to analyze travelers' heterogeneity when evaluating transit service quality (de Oña et al., 2016). Ding and Zhang (2016) applied hierarchical clustering analysis and multinomial logit models to analyze travel mode choice. Depaire et al. (2008) applied LCCA to identify clusters with homogeneous traffic crash patterns and then used multinomial logit to assess the risk factors of each cluster. Chang et al. (2019) and Liu and Fan (2020) also used a two-step method, i.e., LCCA plus mixed logit models, to investigate injury severity in traffic crashes. Prior research has shown that applying such two-step method can reveal hidden relationships and generate richer insights for decision-makers (Chang et al., 2019; de Oña et al., 2016).

The first step of the proposed LSDT is to use LCCA to segment the entire sample into subgroups with similar characteristics. The main reason we are using LCCA here is that it is a probability-based parametric clustering technique, which has been applied in the previous travel behavior literature to identify market segments and has shown its strength in analyzing heterogeneity (e.g., Kim et al., 2019; Wang et al., 2021). In a companion paper of this study, Wang et al. (2021) applied LCCA to residents from low-income neighborhoods in Michigan and they identified three latent clusters based on their current usage of shared modes (including fixed-route public transit and ridehailing services) and their preferences for a proposed MOD transit system; the three clusters include shared-mode enthusiast, shared-mode opponent, and fixed-route transit loyalist. Results indicate varying MOD preferences among the three segments, which intrigues us to further analyze the decision rules regarding MOD preferences in different segments. Therefore, in this paper, we decide to use LCCA to segment people from low-income neighborhoods based on their current transit/ridehailing usage to answer **RQ 1**.

In the second step of the proposed LSDT, we propose to use DT to conduct cluster-specific analysis, instead of using logit models like some previous work did (Chang et al., 2019; Depaire et al., 2008; Liu and Fan, 2020). The main reason is that most logit models have certain limitations due to their predefined assumptions, e.g., the assumption of the independence of irrelevant alternatives [IIA] for multinomial logit models and random parameter distributions for mixed logit models. Once the assumptions are violated, the estimation of the likelihood function will be erroneous (de Oña et al., 2016). In addition, logit models take on the inflexible functional forms to model the relationships between the input and response variables, which may not be accurate or even appropriate when there exist high nonlinearities and/or interactions in the data. By contrast, DT models do not rely on these assumptions and have flexible model structure to capture nonlinearities and interactions. Moreover, DT models offer graphic representation and transparent interpretation for policy making (James et al., 2013). By integrating LCCA and DT, we will be able to extract key insights on what factors are associated with people's preferences for MOD transit and how these relationships vary across different traveler groups determined by their current shared mode usage (**RQ 2**).

[1] To our knowledge, no work has used LCCA's probability estimates as case weights when training cluster-specific DT models.

The reminder of this paper is organized as follows. Section 2 provides a literature review on different models used to assess preference heterogeneity in travel behavior. Section 3 describes the study area and the data. Section 4 discusses the overall modeling framework and introduces the formulation of LCCA and DT. Section 5 presents the results. Section 6 synthesizes the findings, discusses the policy implications, concludes the paper with strengths and limitations of the study, and identifies future research directions.

2 Literature Review

Different individuals would react to the new MOD transit system distinctively due to preference heterogeneity (Bhat, 1997; Fu, 2020). Understanding and analyzing preference heterogeneity can help decision-makers develop better-targeted policies to meet the travel needs of all residents who live in low-income communities.

In the past several decades, mixed logit models have been widely utilized to assess preference heterogeneity (Train, 2009; Yan et al., 2019a). Despite having better model fit than simpler logit models (e.g., multinomial logit and ordered logit models), the mixed logit models have suffered from several drawbacks. Specifically, the mixed logit models rely on the mathematical assumptions about random parameter distributions and error term distributions (Walker and Ben-Akiva, 2002), but these assumptions could easily be violated in real-world applications. In addition, the mixed logit models require extensive work in model tuning and high computational costs. Moreover, some argued that the mixed logit models tend to become quite complex, which makes them less transparent for direct interpretation (Fu, 2020).

Alternative to the mixed logit models, the latent class model (LCM), also known as the latent class choice model, has been developed to study preference heterogeneity (Shen, 2009). The LCM contains two sub-models, i.e., the *class membership model* and the *choice model*. More specifically, the LCM first separates the population into different segments with a class membership model, which maximizes within-segment homogeneity and between-segment heterogeneity; it then estimates segment-specific choice models to reveal the preference heterogeneity residing in the effects of explanatory variables (Kim and Mokhtarian, 2018). The LCM allows researchers to identify various population segments with distinctive preferences, and and it has been wildly applied to assess preference heterogeneity in travel behavior studies (Eldeeb and Mohamed, 2020; Fu, 2020; Kim and Mokhtarian, 2018; Oliva et al., 2018; Shen, 2009; Vij et al., 2013; Wen et al., 2012). For example, Vij et al. (2013) incorporated the influence of latent modal preferences on travel mode choice behavior by using LCM. Recently, Fu (2020) applied LCM to study how a traveler's habit moderates his/her mode choice for commuting trips.

However, the LCM only allows for one dependent variable when conducting the joint estimation for both the class membership model and the choice model, bringing many limitations to real-world applications that may require different dependent variables for the two models and/or need multiple dependent variables (also known as indicators) when conducting clustering analysis. A two-step method (i.e., a clustering step followed by a cluster-specific modeling step) can relax this constraint and

has recently been used to model and interpret people's travel behavior, (e.g., Ding and Zhang, 2016; de Oña et al., 2016). For instance, de Oña et al. (2016) integrated LCCA and DT to assess the perceived transit service quality and detect specific needs and requirements from different subgroups with unique traveler profiles.

3 Study Area and Data

This study investigates heterogeneous traveler preferences for a MOD transit system among low-income neighborhoods. We distributed a web-based survey in the city of Detroit and the city of Ypsilanti area, Michigan, both of which are low-source communities in the region with a significant proportion of the population living under poverty[2]. Participants were recruited from July to November 2018. We obtained a total of 497 and 534 completed responses from Ypsilanti and Detroit, respectively. After removing invalid responses and observations with missing values, a total of 825 (Ypsilanti: 410; Detroit: 415) responses were retained for further analysis. The survey collected data on travelers' stated preferences for MOD transit versus fixed-route system, their current usage of shared mobility, their demographic and socioeconomic characteristics, and built-environment factors. More details of the survey design and distribution can be found in Yan et al. (2019b).

The descriptive statistics of the variables considered in this paper are summarized in Table 1. In the last column of the table, we show in which model(s) the variable is included. Note that *MOD Transit Preference* is the response variable for DT, while *Ridehailing Usage Frequency* and *Bus Usage Frequency* are the indicators for LCCA. Note that as Likert scale (i.e., ordinal) variables with five or more categories can usually be treated as continuous with little concerns (Johnson and Creech, 1983; Norman, 2010; Rhemtulla et al., 2012; Sullivan and Artino Jr, 2013), here, we treat the Likert scale variable (i.e., *MOD Transit Preference*) as a continuous one and apply regression trees to interpret people's preferences for MOD transit across various population groups.

4 Methodology

4.1 Modeling Framework

In this paper, we adopt a two-step latent segmentation based decision tree (LSDT) method, i.e., an integrated approach with LCCA and DT, to assess preference heterogeneity for MOD transit service in low-income communities. Figure 2 illustrates the overall modeling framework.

As shown in Figure 2, the first step is to collect the individual-level travel preference data using survey tools, which will be covered in the next section. Then, LCCA is applied to segment the dataset into K different clusters, each of which represents

[2] According to the American Community Survey 2013-2017 5-year estimates, the median household income in the city of Detroit and the city of Ypsilanti was $27,838 and $35,896 respectively, and the poverty rate was 37.9% and 30.9% respective.

Table 1 Description of variables used in the paper

Variable	Description	%	Mean	SD[3]	Model
Male	0: Female	51.64%			LCCA & DT
	1: Male	48.36%			
Age	1: Under 25	9.45%			LCCA
	2: 25-29	20.61%			
	3: 30-39	32.36%			
	4: 40-49	16.97%			
	5: 50-59	13.21%			
	6: 60-69	6.06%			
	7: 70 or over	1.33%			
Household Income	1: Less than $25,000	25.09%			LCCA
	2: $25,000-$49,999	25.94%			
	3: $50,000-$74,999	16.12%			
	4: $75,000-$99,999	14.18%			
	5: $100,000-$124,999	10.55%			
	6: $125,000-$149,999	5.09%			
	7: $150,000 or more	3.03%			
Car Ownership	0: Do not own any vehicle	21.33%			LCCA & DT
	1: Own at least one vehicle	78.67%			
College Degree	0: Do not have a college degree	51.03%			LCCA & DT
	1: Have a college degree	48.97%			
No Smartphone	0: Have a smartphone	89.09%			LCCA
	1: Do have not a smartphone	10.91%			
No Data Plan	0: Have a data plan	87.27%			LCCA & DT
	1: Do not have a data plan	12.73%			
Ridehailing Experience	0: Never heard of Uber/Lyft or used Uber/Lyft in the past week	38.42%			DT
	1: Used Uber/Lyft at least once in the past week	61.58%			
Within Transit Service Area	0: Do not live within a quarter-mile to a bus stop	59.15%			DT
	1: Live within a quarter-mile to a bus stop	40.85%			
Job Accessibility by Transit	Number of jobs reachable within 45 min of transit travel time		10260.74	19729.92	DT
MOD Transit Preference	1: Strongly prefer fixed-route over MOD	2.91%			DT
	2: Sort of prefer fixed-route over MOD	9.70%			
	3: Not sure	21.33%			
	4: Sort of prefer MOD over fixed-route	32.24%			
	5: Strongly prefer MOD over fixed route	33.82%			
Ridehailing Usage Frequency	0: Never used ridehailing in the past week	38.42%			LCCA
	1: Used ridehailing once in the past week	25.21%			
	2: Used ridehailing twice in the past week	22.30%			
	3: Used ridehailing 3-4 times in the past week	12.24%			
	4: Used ridehailing 5 or more times in the past week	1.82%			
Fixed-Route Transit Usage Frequency	0: Never used bus in the past week	28.24%			LCCA
	1: Used bus once in the past week	9.33%			
	2: Used bus twice in the past week	19.03%			
	3: Used bus 3-4 times in the past week	25.82%			
	4: Used bus 5 or more times in the past week	17.58%			

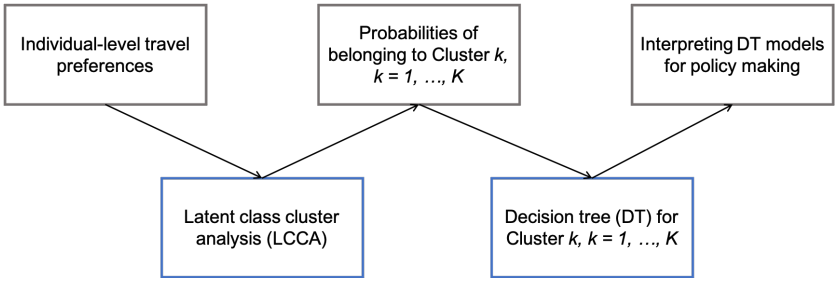

Fig. 1 Overall modeling framework for assessing MOD transit service preference in low-income communities by using LSDT.

distinctive traveler profiles. In particular, we estimate the probabilities of an observation belonging to different latent classes and weight all the observations with the

cluster-specific probabilities when training DT models for different clusters. Compared to directly splitting the dataset into subsets (i.e., deterministic classification), our method (probabilistic classification) enables DT to use the full dataset (i.e., full information) to train three cluster-level DT models, which are distinct from each other due to different weights applied. Moreover, probabilistic classification usually generates more homogeneous results and fewer noises within each cluster, which could lead to a clearer path of the decision rules. These cluster-specific DT models can then allow us to analyze the heterogeneous traveler preference for MOD transit in order to engage more nuanced policy discussions and develop better-targeted policy intervention strategies for low-income neighborhoods.

4.2 Latent Class Cluster Analysis

Latent class cluster analysis (LCCA) is a probabilistic based clustering technique. Figure 2 presents the model framework of the simplified LCCA modified from Wang et al. (2021) [4]. The LCCA model contains two sub-models: The membership model and the measurement model. Specifically, the *membership model* uses active covariates z to predict the latent class membership k, i.e., the latent shared mobility usage segment. In this simplified LCCA model, active covariates include demographic and socioeconomic traits (i.e., gender, age, race, education attainment), travel-related traits (i.e., vehicle ownership), and technology usage (i.e., smartphone and data plan ownership). Note that we retain covariates that relate to job accessibility as inactive covariates, which does not influence the latent class structure. Instead, we will use the retained inactive covariates as inputs for DT to predict the *MOD Transit Preference*. In the *measurement model*, we use the latent variable k to capture the association between the two observed ordinal indicators y: *Ridehailing Usage Frequency* and *Fixed-Route Transit Usage Frequency*. Under the local independence assumption, the two indicators are assumed to be mutually independent given Cluster k.

Following the notation in Vermunt and Magidson (2016), Eqs. (1)–(3) are the mathematical representation of the LCCA model presented in Figure 2. Eq. (1) represents the probability of observing the two indicators y_i for individual i given a set of observed covariates z_i. The unobserved latent class k, which has K categories, intervenes between y_i and z_i. Specifically, $P(y_i|z_i)$ is the probability of the membership model and $P(y_i|k)$ is the probability of the measurement model. Given the local independence assumption, the probability of the measurement model could write as the probability product of the two indicators, i.e., $\prod_{t=1}^{2} P(y_{it}|k)$.

$$P(y_i|z_i) = \sum_{k=1}^{K} P(k|z_i)P(y_i|k) = \sum_{k=1}^{K} P(k|z_i)\prod_{t=1}^{2} P(y_{it}|k) \tag{1}$$

[4] The LCCA model used in this paper is a simplified/constrained version of the one presented in Wang et al. (2021). We modify the LCCA model structure to fit the proposed methodological framework described in Section 2. Major modifications include (1) we remove *MOD Transit Preference* from the indicator set and use it as the output of the DT models; (2) we simplify the covariate set of the full LCCA model. The simplified/constrained LCCA model has consistent results with the full LCCA model regarding variables used in both models.

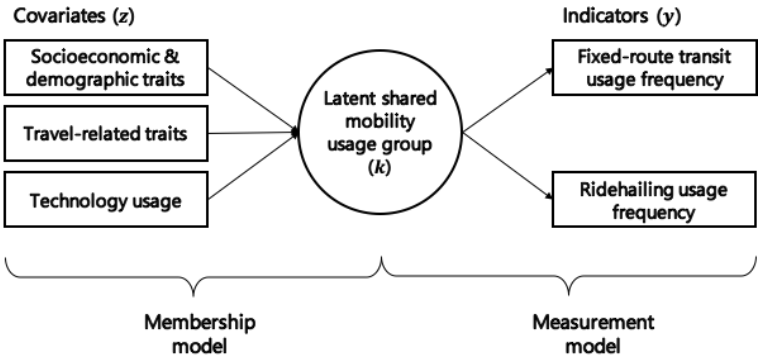

Fig. 2 Model framework of the latent class cluster analysis (LCCA).

Eq. (2) defines the probability of individual i belonging to latent class k given a set of observed covariates z_i, which is parameterized using the multinomial logit formula. For each latent class, we estimate an intercept γ_{k0} and a set of parameters γ_{kr} corresponding to the R active covariates.

$$P(k|z_i) = \frac{exp(\gamma_{k0} + \sum_{r=1}^{R} \gamma_{kr} z_{ir})}{\sum_{k'=1}^{K} exp(\gamma_{k'0} + \sum_{r=1}^{R} \gamma_{k'r} z_{ir})} \tag{2}$$

Eq. (3) defines the probability of individual i with its tth indicator equal to m given the latent class k. Note that both indicators used in this study are ordinal variables. As such, the probability is parameterized using the adjacent-category logit formula. We estimate an intercept for each ordinal value m and a parameter β_k^t for each latent class. Here, the $y_m^{t^*}$ is the score assigned to level m of the tth indicator.

$$P(y_{it} = m|k) = \frac{exp(\beta_m^t + \beta_k^t \cdot y_m^{t^*})}{\sum_{m'=1}^{M_t} exp(\beta_{m'}^t + \beta_k^t \cdot y_{m'}^{t^*})} \tag{3}$$

In this paper, we estimate the LCCA model by using Latent GOLD software (v.5.1). Three clusters are achieved from our analysis, and the detailed results are covered in Subsect. 5.1.

4.3 Decision Trees

Decision trees (DT) can be used to tackle both regression and classification problems, and in this paper, we treat the five-level *MOD Transit Preference* variable as continuous and fit regression trees to explain the heterogeneity in people's travel preferences. DT can automatically capture complex high-dimensional data and is famous for its

intelligible graphical representation and transparent interpretation. Despite many different methods to fit DT, the classification and regression trees (CART) algorithm is probably the most popular one for tree induction (Breiman et al., 1984). The following description is focused on regression part of CART.

DT recursively partitions the feature space into sub-regions until some stopping rule is applied (Hastie et al., 2009). Suppose we have the data with each observation denoted by (x_i, y_i) and its case weight w_i, we consider a splitting variable j and split point s; then, the pair of half-planes are defined as:

$$R_1(j,s) = \{X|X_j \leq s\}, R_2(j,s) = \{X|X_j > s\}. \tag{4}$$

Then, we aim to estimate the splitting variable j and split point s by solving

$$\min_{j,s} \Big\{ \min_{c_1} \sum_{x_i \in R_1(j,s)} w_i(y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} w_i(y_i - c_2)^2 \Big\}. \tag{5}$$

For any j and s, the inner minimization is achieved by $\hat{c}_1 = \text{ave}(y_i|x_i \in R_1(j,s))$ and $\hat{c}_2 = \text{ave}(y_i|x_i \in R_2(j,s))$, where ave$(\cdot)$ indicates a weighted average function.

After finding the best split, we can partition the data into two regions and repeat the partition process until a stopping criterion is reached. Such a large tree can be denoted by T_0. However, a very large tree tends to overfit the data, so we need to control the tree size to achieve the best out-of-sample performance. Therefore, the tree is often pruned by using *cost-complexity pruning* (Hastie et al., 2009). The cost complexity criterion is

$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{x_i \in R_m} w_i(y_i - \hat{c}_m)^2 + \alpha|T| \tag{6}$$

where $|T|$ is the number of terminal nodes (leaves) in tree T, $\hat{c}_m = \text{ave}(y_i|x_i \in R_m)$, and α is the complexity parameter. Here, we aim to find, for each $\alpha \geq 0$, the subtree $T_\alpha \subseteq T_0$ to minimize $C_\alpha(T)$. There is clearly a trade-off between tree size and its goodness-of-fit to the data. We can select a value of α by using cross-validation, and then return to the entire dataset to output the subtree corresponding to α.

A key output of DT is variable importance, which assesses the impacts of independent variables on the DT model's prediction. In our case of regression trees, variable importance is estimated by the decrease in node impurities from splitting on the variable, where node impurity is measured by residual sum of squares.

In this paper, we apply the CART algorithm by using the R package *rpart* Therneau et al. (2015) and the trees are visualized using *rpart.plot* Milborrow (2020). We use grid search to tune the main hyperparameters of DT models, including *minsplit* (the minimum number of observations that must exist in a node in order for a split to be attempted), *maxdepth* (the maximum depth of any node of the final tree, with

the root node counted as depth 0), and cp (complexity parameter). For the benchmark DT model (trained on the overall sample), $minsplit = 20, maxdepth = 6$, and $cp = 0.013$. For the DT model built for Cluster 1, $minsplit = 14, maxdepth = 4$, and $cp = 0.017$; for Cluster 2, $minsplit = 18, maxdepth = 3$, and $cp = 0.010$; for Cluster 3, $minsplit = 18, maxdepth = 10$, and $cp = 0.014$.

5 Results

5.1 Latent Class Cluster Analysis

To select the optimal number of latent classes, we run the LCCA model with varying numbers of clusters from 1 to 10. The Bayesian Information Criterion, or BIC (equals to 4561.57), indicates the 3-cluster solution has the best model fit after penalizing model complexities; the solution also has a good interpretability. As such, we choose the 3-cluster LCCA as the final model. Table 2 presents parameters and z-values of both membership and measurement models of the 3-cluster LCCA solution. We name and develop cluster profiles based on the cluster-specific distributions of the indicators and covariates (see Table 3).

As shown in Table 3, **Cluster 1** is the largest cluster among the three, which comprises 50% of the full sample. Cluster 1 members have an average ridehailing usage frequency of 2.03, indicating a more-than-twice usage of ridehailing services in the past week, which is the highest among the three clusters. Meanwhile, Cluster 1 members also have a relatively high fixed-route transit usage frequency (2.44). As such, we name Cluster 1 as "**shared-mode user**." The **shared-mode user** cluster comprises a slightly larger proportion of males than the sample average (53% versus 48%). Among the three clusters, shared-mode users have the largest proportion of individuals who are younger than 40 years old (71%) and own college degrees (64%). They also have the highest household income. A large proportion of individuals from this cluster own a vehicle (88%), whereas 11% and 15% individuals do not have a smartphone or data plans, respectively.

Cluster 2 comprises 29% of respondents in the sample. Their average ridehailing and fixed-route transit usage frequencies are 0.26 and 0.39, which are the lowest among the three clusters, respectively. Reflective of their low usage of shared modes, we name Cluster 2 as "**shared-mode non-user**." The **shared-mode non-user** cluster contains more females than males (64% versus 36%). More than half of the individuals in this cluster have a college degree (54%). Moreover, shared-mode non-users have the highest proportion of vehicle owners (94%), smartphone owners (97%), and data plan owners (97%) among the three clusters.

Cluster 3 comprises 21% of respondents in the sample. Cluster 3 members have the lowest usage of ridehailing services (0.20) and the highest fixed-route transit usage frequency (2.98) among the three clusters. Thus, we name Cluster 3, "**transit-only user**." Compared to the other two clusters, the **transit-only user** cluster has the largest proportion of elderly people (60 years and above, 17%), and the largest proportion of the low-income group (63% of the individuals have a household income less than $25,000). Only 5% of individuals from the **transit-only user** cluster have

Table 2 LCCA model coefficients

Measurement model						
β_k^t	**Cluster 1** *Shared-mode user*		**Cluster 2** *Shared-mode non-user*		**Cluster 3** *Transit-only user*	
	Coefficient	z-value	Coefficient	z-value	Coefficient	z-value
Ridehailing usage frequency	**3.44**	3.82	**-1.57**	-3.31	**-1.87**	-3.87
Fixed-route transit usage frequency	**0.36**	4.79	**-1.14**	-9.40	**0.78**	7.78
β_m^t	**Ridehailing usage frequency**		**Fixed-route transit usage frequency**			
	Coefficient	z-value	Coefficient	z-value		
0: Never used in the past week	**4.08**	5.35	0.25	1.72		
1: Used once in the past week	**4.56**	3.89	**-0.33**	-2.35		
2: Used twice in the past week	**1.60**	5.43	**0.35**	4.18		
3: Used 3-4 times in the past week	**-2.44**	-3.77	**0.30**	2.87		
4: Used 5 or more times in the past week	**-7.79**	-5.02	**-0.57**	-3.70		
Membership model						
γ_{k0}, γ_{kr}	**Cluster 1** *Shared-mode user*		**Cluster 2** *Shared-mode non-user*		**Cluster 3** *Transit-only user*	
	Coefficient	z-value	Coefficient	z-value	Coefficient	z-value
Intercept	**1.54**	5.42	**-1.46**	-3.58	-0.08	-0.18
Gender						
Female	-0.0073	-0.10	**0.35**	4.14	**-0.35**	-2.97
Male	0.0073	0.10	**-0.35**	-4.14	**0.35**	2.97
Age	**-0.26**	-5.41	0.04	0.73	**0.22**	3.03
Income	**0.22**	4.09	0.05	0.86	**-0.28**	-2.85
College degree						
Yes	**0.53**	4.97	**0.31**	2.62	**-0.84**	-4.27
No	**-0.53**	-4.97	**-0.31**	-2.62	**0.84**	4.27
Car ownership						
Yes	0.10	0.98	**0.73**	5.11	**-0.84**	-6.60
No	-0.10	-0.98	**-0.73**	-5.11	**0.84**	6.60
Smart phone						
Have a smart phone	**-0.28**	-2.25	**0.53**	2.72	-0.25	-1.37
Do not have a smart phone	**0.28**	2.25	**-0.53**	-2.72	0.25	1.37
Mobile data plan						
Have a mobile data plan	**-0.23**	-2.02	**0.69**	3.56	**-0.45**	-2.76
Do not have a data plan	**0.23**	2.02	**-0.69**	-3.56	**0.45**	2.76

Notes: Coefficients use effect coding. The bolded coefficients are statistically significant at the 0.05 level.

college degrees and only 32% own vehicles, which are much lower than the counterparts of the other two clusters. The **transit-only user** cluster also has the highest proportions of individuals who do not have smartphones (21%) or data plans (23%) among all three clusters.

5.2 Decision Trees

As illustrated in Figures 3-6, four different regression trees have been generated. Specifically, Figure 3 is the DT for overall sample of travelers; Figures 4-6 correspond to each of the detailed traveler profiles of three different clusters. As illustrated

Table 3 Cluster-specific shares/means of covariates

	Cluster 1 *Shared-mode user*	**Cluster 2** *Shared-mode non-user*	**Cluster 3** *Transit-only user*	***Sample***
Cluster size	0.50	0.29	0.21	*1.00*
ridehailing usage frequency	**2.03**	0.26	0.20	*1.14*
Fixed-route transit usage frequency	2.44	0.39	**2.98**	*1.95*
Gender				
Female	0.47	**0.64**	0.46	*0.52*
Male	0.53	0.36	**0.54**	*0.48*
Age				
1: Under 25	**0.11**	0.06	**0.11**	*0.09*
2: 25-29	**0.26**	0.17	0.14	*0.21*
3: 30-39	0.34	**0.40**	0.17	*0.32*
4: 40-49	0.17	0.16	**0.18**	*0.17*
5: 50-59	0.11	0.11	**0.22**	*0.13*
6: 60-69	0.02	0.06	**0.16**	*0.06*
7: 70 or over	0.00	**0.03**	0.01	*0.01*
Household income				
1: Less than $25,000	0.11	0.22	**0.63**	*0.25*
2: $25,000-$49,999	**0.27**	0.24	0.26	*0.26*
3: $50,000-$74,999	0.19	**0.20**	0.04	*0.16*
4: $75,000-$99,999	0.16	**0.17**	0.06	*0.14*
5: $100,000-$124,999	**0.14**	0.10	0.01	*0.11*
6: $125,000-$149,999	**0.08**	0.04	0.00	*0.05*
7: $150,000 or more	**0.05**	0.02	0.00	*0.03*
Has a college degree	**0.64**	0.54	0.05	*0.49*
Vehicle owner	0.88	**0.94**	0.32	*0.79*
Do not own smartphone	0.11	0.03	**0.21**	*0.11*
Do not have data plan	0.15	0.03	**0.23**	*0.13*

Note: The numbers in this table represent expected values for the segment, computed with posterior class membership probabilities. Bolded numbers are the maximum segment-specific shares/means across the three segments, whereas the underlined numbers are the minimum segment-specific shares/means across the three segments. See Table 1 for variable definitions.

in Table 1, *MOD Transit Preference* is selected as the response variable, while seven other variables are chosen as the independent variables. The selection of independent variables is mainly based upon the results from Yan et al. (2019b), which found these seven variables are statistically significant when used to model people's stated preferences for the MOD transit service.

We use mean absolute error (MAE) to measure the performance of the DT models. MAE is formally defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{7}$$

where $\hat{y}_i$ is the predicted value for observation i, y_i is the true value for observation i, and n is the number of observations in the testing set. The overall MAE estimate from the joint DT models for K clusters can be computed as

$$\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{k=1}^{K}\hat{y}_{i,k}p_{i,k}-y_i\right| \tag{8}$$

where $\hat{y}_{i,k}$ is the predicted value for observation i from the DT model for Cluster k, and $p_{i,k}$ is the probability that observation i belongs to Cluster k with $\sum_{k=1}^{K} p_{i,k} = 1$. By using leave-one-out cross-validation and Eqs. (7) and (8), we estimate the MAE of the DT model for the overall sample is 0.833, while the overall MAE from the joint DT models for the three clusters is 0.829. Hence, we find that by applying the proposed framework illustrated in Figure 2, the LSDT method shows similar (or, even slightly better) predictive accuracy than the benchmark DT model.

For the fitted trees (see Figures 3-6), each box denotes a tree node, and the nodes at the bottom are called terminal nodes. In each node, we indicate the total number of observations belonging to this node, the corresponding percentage of observations in the node, and the average value (i.e., the fitted value) of the dependent variable (i.e., *MOD Transit Preference*) among all the observations in this node. The coloring of the node boxes are based on the fitted value: Darker the blue, larger the fitted value. Under each node, the left branch indicates 'yes' to the condition listed there, while the right branch denotes 'no' to the condition.

In Figure 3, we show the DT built for the overall sample of travelers. The primary split for the overall sample is based on *Job Accessibility by Transit*, which is the same case for the three cluster-specific DT models. An important insight we gain here is when job accessibility is very high (above 52k), travelers are in general more favorable of fixed-route transit service; when job accessibility is below 52k, people are more open to MOD transit, but have much more complex decision rules. For example, Node 13 indicates that with job accessibility less than 4,025 (much lower than the mean of job accessibility, i.e., 10,261), having previous ridehailing experience, and owning a college degree, these travelers are very supportive of MOD transit (with the fitted value of 4.4), consisting of 22% of the overall sample. Therefore, we may conclude that people who have high job accessibility could go to work easily by using the existing fixed-route transit services. In other words, fixed-route transit may have already met their travel demands; as such, they do not necessarily need MOD transit. In contrast, MOD transit can serve as an affordable alternative for people who are currently having low job accessibility.

From Figure 4 to 6, we show the DT models for the three latent clusters. These three DT models have the entire sample as the input data (i.e., $n = 825$ at the top node), but different case weights (estimated from LCCA to represent the individual's probabilities of belonging to each cluster) are applied when fitting models for different clusters. Note that, for some nodes, we may observe that the number of observations seem inconsistent with the percentage of observations in the node: Taking the

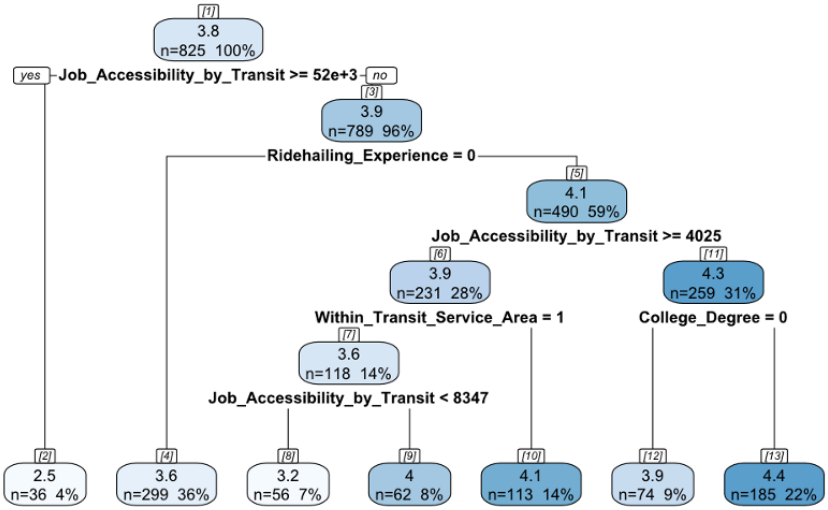

Fig. 3 Decision tree for overall sample of travelers.

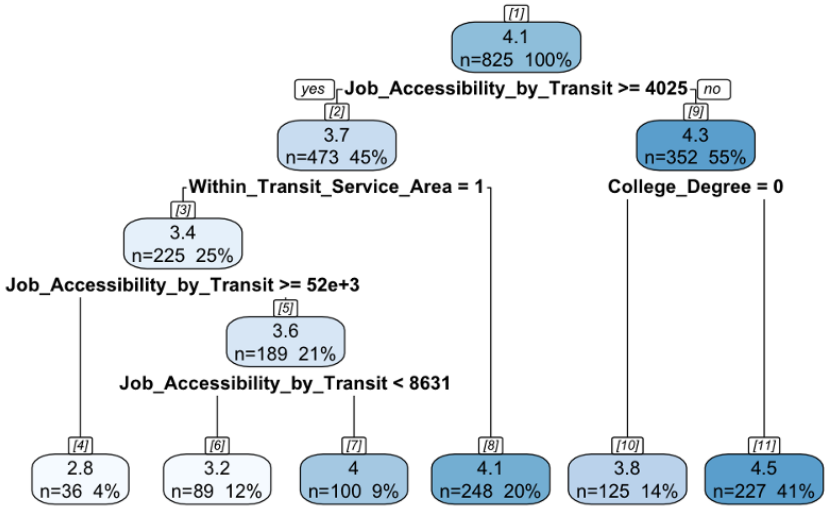

Fig. 4 Decision tree for Cluster 1 "Shared-mode user."

DT for Cluster 1 as an example (see Figure 4), Node 2 consists of 473 observations and 45% of the sample, while Node 9 has 352 observations and 55% of the sample. This is because the percentage shown here is a weighted percentage using the case weights (i.e., probabilities of belonging to different clusters) passed to the CART al-

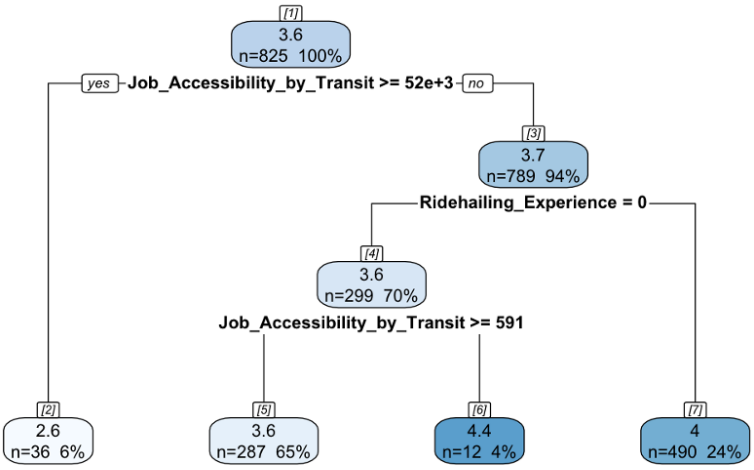

Fig. 5 Decision tree for Cluster 2 "Shared-mode non-user."

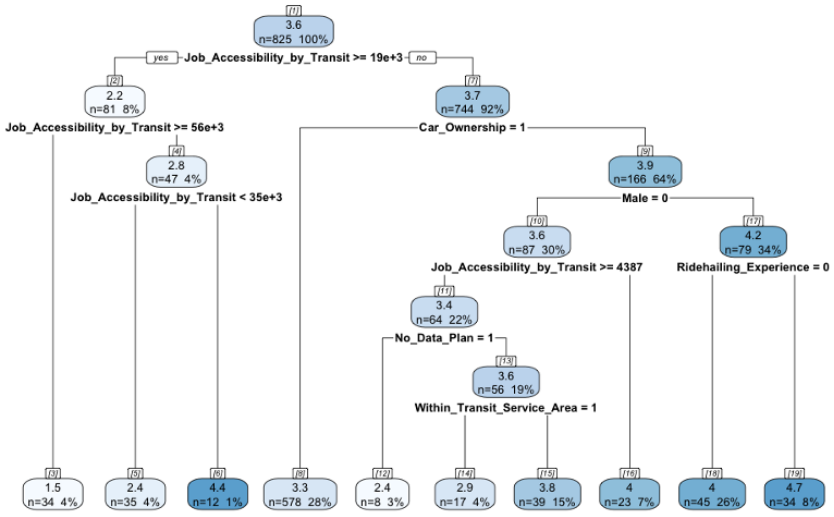

Fig. 6 Decision tree for Cluster 3 "Transit-only user."

gorithm (Milborrow, 2020). Therefore, when interpreting the DT models, we mainly focus on the percentage of observations instead of the absolute observation counts.

Figure 4 illustrates the DT built for Cluster 1 "Shared-mode user." This tree is similar to the DT for the overall sample, but *Ridehailing Experience* is not included in this tree. In addition, Node 1 (i.e., root node) of Cluster 1 model has the fitted value

of 4.1, which is larger than the fitted values for the root nodes of the other two cluster-specific DT models. These results suggest that shared-mode users are more open to different shared modes and have higher preferences for MOD transit. In addition, the DT for Cluster 1 uses *College Degree* for splitting, which is not included in the other two cluster-specific DT models. For shared-mode users who have a college degree and low job accessibility (i.e., Node 11), they are very supportive of MOD transit. This is consistent with the existing findings that travelers who are more highly educated are more open to new mobility options (Lavieri and Bhat, 2019). Moreover, this DT model also shows that shared-mode users who have better job accessibility but living outside the transit service area are more willing to adopt MOD transit. This finding indicates the potential of MOD transit to tackle the infamous first-/last-mile problem in the U.S.

Figure 5 shows the DT built for Cluster 2 "Shared-mode non-user." This tree looks much simpler compared to the DT models for the overall sample and the other two clusters. Only two variables, namely, *Job Accessibility Transit* and *Ridehailing Experience*, are included in the model. As shown in Node 5, the majority (i.e., 65% of the sample) of the shared-mode non-users are approximately neutral when comparing the fixed-route with MOD transit.

Figure 6 represents the DT built for Cluster 3 "Transit-only user." This tree is the most complicated one among the three cluster-specific DT models. Six different variables show up in this tree, in comparison to four included in the overall sample tree, three in Cluster 1 tree, and two in Cluster 2 tree. An important observation is that with relatively higher job accessibility (more than 19k), transit-only users have higher preference for fixed-route over MOD transit. In contrast, according to the other two cluster-specific DT models, the threshold of job accessibility is much higher for choosing fixed-route over MOD transit (i.e., the fitted values of *MOD Transit Preference* less than 3): 52k for shared-mode users who also live within transit service area (Node 4 in Figure 4) and 52k for shared-mode non-users (Node 2 in Figure 5). Compared to other two types of travelers who would choose fixed-route only if the job accessibility is exceptional, transit-only users tend to stick to fixed-route transit, when the job accessibility is acceptable. However, for the DT model built for the overall sample, the job accessibility threshold is 52k (Node 2 in Figure 3), which demonstrates that the proposed LSDT method can generate rather richer insights than a single DT could. We also find that with relatively lower job accessibility (less than 19k), transit-only users who have access to personal vehicles have relatively lower preference for MOD transit than the ones who have no access to personal vehicles do. But the difference is small, i.e., 3.3 for car owners versus 3.9 for carless people. Among those carless transit-only users, male travelers are more acceptive of MOD transit than females. This observation is consistent with the results in Yan et al. (2019b), which also finds that females might have safety concerns regarding the new MOD transit service. Besides, among those female transit-only users, despite acceptable job accessibility, no data plan could lead to low acceptance of MOD transit (Node 12), which shows the importance of addressing digital divide when deploying the new MOD transit system.

There exist several seemingly unreasonable nodes in the trees, i.e., Nodes 8 and 9 in Figure 3, Nodes 6 and 7 in Figure 4, and Nodes 5 and 6 in Figure 6. These nodes all

have a same problem that with job accessibility below certain thresholds, travelers are less likely to choose MOD transit. As the DT models are purely data-driven without relying on any predefined assumptions, these anomalies are usually caused by the noise/bias in the data and overfitting of the DT models.

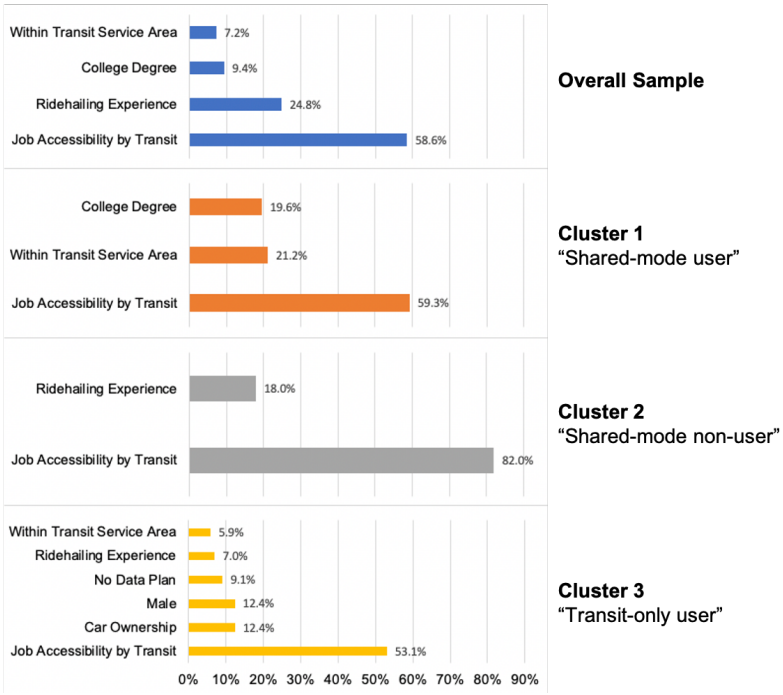

Fig. 7 Relative variable importance plots for four DT models

In Figure 7, we present the relative variable importance plots (scaled to sum up to 100% for each DT model) for the four DT models. We find that *Job Accessibility by Transit* is the most important variable for all four DT models. Thus, it seems that traveler preferences are mostly shaped by the destinations accessible via transit. This finding is consistent with the notion that accessibility—rather than mobility—represents people's basic need for transportation (Levine et al., 2019). On the other hand, whether living within or outside the transit service area shapes the preferences of "shared-mode users" (Cluster 1), indicating the importance of last-mile transit connectivity. *Ridehailing Experience* is the second-most important variable for the overall sample tree and Cluster 2 tree. This indicates that for shared-mode non-users, having used ridehailing at least once in the past week is an important indicator to gauge traveler preferences for MOD transit. Moreover, *Car Ownership* and *Male* are important variables in the decision tree model for Cluster 3 ("Transit-only user"), but they are insignificant in the other models. According to the population profiles shown in Table 3, the vast majority of individuals in Cluster 1 and Cluster 2 have access to a data plan and own a personal vehicle. The lack of variability may explain why

they are not important predictors of MOD transit preference in all models expect the Cluster 3 model. This finding further verifies the importance of fitting cluster-specific models, as an all-sample model may suppress the heterogeneous preferences across population segments. Interestingly, a lower preference for MOD transit exists among females in Cluster 3 but not those in the other two clusters. A possible explanation is that some females in Cluster 3 might have unpleasant experiences with or negative perceptions of ridehailing.

6 Discussion and Conclusion

According to the results presented in the previous section, we find that the LSDT method can generate much richer insights than the a single DT model fitted for the overall sample. In particular, when combining the results from LCCA and cluster-specific DT, we can attach the traveler class profile to their corresponding decision rules when choosing between MOD transit and fixed-route services.

For example, the LCCA results for Cluster 3 suggest that the travelers in this cluster are most vulnerable (i.e., having the largest proportion of older, low-income, carless, and technology illiterate people than the other two clusters) yet most dependent on public transit services. When investigating their decision rules shown in Figure 6, we find that for people who are currently enjoying very high job accessibility by transit want to stick with the fixed-route services and those people are very likely to live in the downtown area (Yan et al., 2019b), so we may want to keep running the fixed-route service in the downtown region especially between major corridors. On the other hand, for people have relatively lower accessibility and no access to personal vehicles, females are more reluctant to choose MOD transit than males due to safety concerns (Yan et al., 2019b), so to successfully serve low-income neighborhoods located in lower-density areas, we need to come up with innovative strategies to improve the safety of on-demand shuttles. Some solutions include instead of sending travelers to their doorsteps, the on-demand shuttles would send them to a virtual stop that is located in the central area of the community in order to reduce the concerns from female travelers.

In contrast, the LCCA results show that the travelers in Cluster 1 have the largest proportion of individuals who are technology savvy, younger than 40, own college degrees, have high household income, and own a vehicle, and they are using public transit and ridehailing service frequently. The DT model for Cluster 1 shows in general this group of people is supportive of the MOD transit service, with only 4% of them are somewhat inclined to the fixed-route transit (Node 4 in Figure 4). Hence, when MOD transit starts to operate, we probably will not lose much transit ridership among this small population group, whereas for the rest majority, they are likely to use the MOD transit to substitute their fixed-route transit trips and potentially, some of their ridehailing trips.

According to the LCCA results for Cluster 2, we find that the travelers in this group have the highest proportion of owning personal vehicles, smartphones, and data plans, and they do not use public transit or ridehailing much in their daily life. The decision rules of Cluster 2 are quite simple and show these people are generally

neutral to MOD transit, with one exception that for the individuals currently have very low job accessibility (Node 6 in Figure 5), they show high potential to adopt MOD transit in the future. Therefore, one insight is that when designing the MOD transit system, we need to expand the service area of the existing transit system and provide on-demand shuttles to fill the transit gaps created by the existing fixed-route services.

To summarize, the insights gained here can help transit agencies and transportation planners and engineers to design an inclusive MOD transit system with higher efficiency and effectiveness. They can also leverage our research findings to develop better-targeted strategies to promote MOD transit usage in low-income communities.

There are some limitations of this study. First, there exists some sampling bias when collecting data in Ypsilanti, Michigan. Unlike Detroit data collection, we did not have in-person recruitment in Ypsilanti, so some low-income population was under-represented in our sample. Second, DT models may sometimes be sensitive to small perturbations, which would lead to unstable model structures.

Future work should increase in-person recruitment among low-income communities to have a less biased sample for analysis. In addition, a model distillation approach could be considered to generate more stable DT models for interpretation (Zhou et al., 2018). Lastly, we want to emphasize that we do not advocate *fully* relying on the proposed method to make policy intervention decisions; instead, we suggest comparing the outputs from different approaches (i.e., our proposed method, logit models, and machine-learning methods) to generate more comprehensive results and insights for decision-making (Zhao et al., 2020).

Acknowledgements

This work was partially supported by Poverty Solutions at the University of Michigan and the U.S. Department of Transportation through the Southeastern Transportation Research, Innovation, Development and Education (STRIDE) Region 4 University Transportation Center (Grant No. 69A3551747104). The previous version of this paper was presented in the 2021 TRB Annual Meeting.

Authors Contributions

The authors confirm contribution to the paper as follows: study conception and design: Zhao, Wang, Yan; data collection: Yan, Zhao; analysis and interpretation of results: Zhao, Wang, Cao, with inputs from Yan; initial draft manuscript preparation: Zhao, Wang, with editing from remaining authors. All authors reviewed the results and approved the final version of the manuscript.

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Bhat CR (1997) An endogenous segmentation mode choice model with an application to intercity travel. Transportation science 31(1):34–48

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC press

Chang F, Xu P, Zhou H, Chan AH, Huang H (2019) Investigating injury severities of motorcycle riders: A two-step method integrating latent class cluster analysis and random parameters logit model. Accident Analysis & Prevention 131:316–326

Depaire B, Wets G, Vanhoof K (2008) Traffic accident segmentation by means of latent class clustering. Accident Analysis & Prevention 40(4):1257–1266

Ding L, Zhang N (2016) A travel mode choice model using individual grouping based on cluster analysis. Procedia engineering 137:786–795

Eldeeb G, Mohamed M (2020) Quantifying preference heterogeneity in transit service desired quality using a latent class choice model. Transportation Research Part A: Policy and Practice 139:119–133

Fu X (2020) How habit moderates the commute mode decision process: integration of the theory of planned behavior and latent class choice model. Transportation pp 1–27

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer

Johnson DR, Creech JC (1983) Ordinal measures in multiple indicator models: A simulation study of categorization error. American Sociological Review pp 398–407

Kim SH, Mokhtarian PL (2018) Taste heterogeneity as an alternative form of endogeneity bias: Investigating the attitude-moderated effects of built environment and socio-demographics on vehicle ownership using latent class modeling. Transportation Research Part A: Policy and Practice 116:130–150

Kim SH, Circella G, Mokhtarian PL (2019) Identifying latent mode-use propensity segments in an all-av era. Transportation Research Part A: Policy and Practice 130:192–207

Lavieri PS, Bhat CR (2019) Investigating objective and subjective factors influencing the adoption, frequency, and characteristics of ride-hailing trips. Transportation Research Part C: Emerging Technologies 105:100–125

Levine J, Grengs J, Merlin LA (2019) From Mobility to Accessibility: Transforming Urban Transportation and Land-Use Planning. Cornell University Press

Liu P, Fan W (2020) Exploring injury severity in head-on crashes using latent class clustering analysis and mixed logit model: A case study of north carolina. Accident Analysis & Prevention 135:105388

Maheo A, Kilby P, Van Hentenryck P (2019) Benders decomposition for the design of a hub and shuttle public transit system. Transportation Science 53(1):77–88

Milborrow S (2020) Plot 'rpart' models: An enhanced version of 'plot.rpart'. R Package 'rpartplot'

Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. Advances in health sciences education 15(5):625–632

Oliva I, Galilea P, Hurtubia R (2018) Identifying cycling-inducing neighborhoods: A latent class approach. International journal of sustainable transportation 12(10):701–713

de Oña J, de Oña R, López G (2016) Transit service quality analysis using cluster analysis and decision trees: a step forward to personalized marketing in public transportation. Transportation 43(5):725–747

Pew Research Center (2018) Mobile fact sheet. Pew Research Center Survey Results Accessed on December 29, 2018 from http://wwwpewinternetorg/fact-sheet/mobile/

Rhemtulla M, Brosseau-Liard PÉ, Savalei V (2012) When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. Psychological methods 17(3):354

Shen J (2009) Latent class model or mixed logit model? a comparison by transport mode choice data. Applied Economics 41(22):2915–2924

Shen Y, Zhang H, Zhao J (2018) Integrating shared autonomous vehicle in public transportation system: A supply-side simulation of the first-mile service in singapore. Transportation Research Part A: Policy and Practice 113:125–136

Sullivan GM, Artino Jr AR (2013) Analyzing and interpreting data from likert-type scales. Journal of graduate medical education 5(4):541–542

Therneau T, Atkinson B, Ripley B, Ripley MB (2015) Package 'rpart'. Available online: cran ma ic ac uk/web/packages/rpart/rpart pdf (accessed on 20 April 2016)

Train KE (2009) Discrete choice methods with simulation. Cambridge university press

Vermunt JK, Magidson J (2016) Technical guide for latent gold 5.1: Basic, advanced, and syntax. Belmont, MA: Statistical Innovations Inc

Vij A, Carrel A, Walker JL (2013) Incorporating the influence of latent modal preferences on travel mode choice behavior. Transportation Research Part A: Policy and Practice 54:164–178

Walker J, Ben-Akiva M (2002) Generalized random utility model. Mathematical social sciences 43(3):303–343

Wang X, Yan X, Zhao X, Cao Z (2021) Identifying latent shared-mode preference segments in low-income communities: Ride-hailing, bus, and mobility-on-demand transit. Proceedings of Transportation Research Board 100th Annual Meeting

Wen CH, Wang WC, Fu C (2012) Latent class nested logit model for analyzing high-speed rail access mode choice. Transportation Research Part E: Logistics and Transportation Review 48(2):545–554

Yan X, Levine J, Zhao X (2019a) Integrating ridesourcing services with public transit: An evaluation of traveler responses combining revealed and stated preference data. Transportation Research Part C: Emerging Technologies 105:683–696

Yan X, Zhao X, Han Y, Van Hentenryck P, Dillahunt T (2019b) Mobility-on-demand versus fixed-route transit systems: an evaluation of traveler preferences in low-income communities. arXiv preprint arXiv:190107607

Zhao X, Yan X, Yu A, Van Hentenryck P (2020) Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. Travel behaviour and society 20:22–35

Zhou Y, Zhou Z, Hooker G (2018) Approximation trees: Statistical stability in model distillation. arXiv preprint arXiv:180807573