A Cluster-based Joint Model of Travel Mode and Departure Time Choices

Ramin Shabanpour (Corresponding Author)
Ph.D. Candidate
Department of Civil and Materials Engineering
University of Illinois at Chicago
842 W. Taylor Street, Chicago, IL 60607-7023
Tel: (312) 996-3441
Email: rshaba4@uic.edu

Nima Golshani
Ph.D. Student
Department of Civil and Materials Engineering
University of Illinois at Chicago
Tel: (312) 996-3441
Email: ngolsh2@uic.edu

Sybil Derrible, Ph.D.
Assistant Professor
Department of Civil and Materials Engineering
University of Illinois at Chicago
Tel: (312) 996-2429
Email: derrible@uic.edu

Abolfazl (Kouros) Mohammadian, Ph.D.
Professor
Department of Civil and Materials Engineering
University of Illinois at Chicago
Tel: (312) 996-9840
Email: kouros@uic.edu

Mohammad Miralinaghi
Ph.D. Candidate
School of Civil Engineering
Purdue University
W. Lafayette, IN 47907, USA
Email: miralinaghi@purdue.edu

Initial submission: 07/26/2016
Revised submission: 11/15/2016
Paper is submitted for presentation at the 96th Annual Meeting of the Transportation Research Board (TRB) and potential publication in Transportation Research Record

A Cluster-based Joint Model of Travel Mode and Departure Time Choices

ABSTRACT

This paper presents a cluster-based joint modeling approach to investigate heterogeneous travelers' behavior toward trip mode and departure time choices by considering them as a joint decision. To do so, we first use a two-step clustering algorithm to classify travelers into six distinct clusters to account for the heterogeneity in their decision-making behavior. Then, we provide a copula-based joint discrete-continuous model for each cluster in which the travel mode and departure time are estimated by a multinomial logit and a log-linear regression model, respectively. These two models are jointly estimated using copula approach. To investigate the performance of the proposed approach, we compare its results with three other models: (i) an aggregate joint model on all non-clustered observations to assess the potential benefits of population clustering, (ii) a cluster-based univariate mode choice model, and (iii) a cluster-based univariate departure time model to analyze the potential advantages of joint modeling technique. The goodness-of-fit measures and prediction accuracy results in all cases demonstrate that the proposed cluster-based joint model significantly outperforms the other models. Furthermore, the variations in the estimated parameters of different clusters indicate significant behavioral differences across clusters. Hence, the proposed cluster-based joint model, while offering higher accuracy, possesses a significant potential for transportation policy-making because it has the capability to target different types of travelers based on their decision-making behavior.

Keywords: Copula-based model, cluster analysis, principal component analysis, mode choice, departure time.

1. INTRODUCTION

Travel mode and departure time choices are two key components of travel behavior that directly influence the spatial and temporal distribution of travel demand in a transportation system (*1*). Recognizing the key influencing factors in travelers' decision behavior, and especially in travel mode and departure time choices, is essential to devise effective transportation demand management (TDM) policies (*2*). These travel choices are closely intertwined because while the time-of-day choice substantially affects the attributes of travel modes (e.g. availability of travel mode, travel time, travel cost), the choice of departure time is also essentially influenced by the expected travel times for each mode of travel. The first issue in modeling these decisions is the shared factors affecting them and/or the causal effects that they have on each other (*3, 4*). Hence, it is necessary to investigate these two travel decisions jointly to capture the unrestricted correlation between their unobserved influencing factors. Numerous studies have been conducted to consider multiple travel dimensions in a joint model structure (*5-7*).

Another critical issue in modeling these interrelated decisions is the heterogeneity of travelers' in terms of decision-making criteria related to various attributes of their available choices. Several studies show that lifestyle has a direct impact on travel behavior, and so, heterogeneous travelers naturally respond differently to different TDM policies, greatly affecting their effectiveness (*8-10*). Therefore, capturing the heterogeneity of travelers offers substantial benefits to develop a model that is both reliable and policy-sensitive in practice.

This study aims to bring together these two lines of research by coupling clustering analysis approach with joint trip departure time and mode choice modeling. To this end, we first conduct clustering analysis technique using a two-step clustering algorithm to assign individuals to certain clusters, in which members of each cluster are relatively homogenous in terms of their lifestyle specifications and decision behavior. Then, we estimate a joint discrete-continuous model to investigate the joint travel decision behavior of departure time and mode choices within each cluster. Travelers' mode choice and continuous departure time choice are estimated by a multinomial logit and a log-linear regression model, respectively. These two models are jointly estimated using copula approach.

The copula-based approach introduced by Bhat and Eluru (*11*) facilitates model estimation without imposing restrictive distribution assumptions on the dependency structures between the errors in the discrete and continuous model components. In this line of research, Eluru et al. (*12*) used copula-based approach to simultaneously estimate vehicle ownership, residential location, and vehicle miles traveled while capturing the correlation between the error terms. Born et al. (*13*) estimated a copula-based joint continuous-discrete model of activity and accompany type and activity duration. Results of these studies show significant improvement of copula-based models over the other models, which ignore the interrelated distribution of the variables. Other examples of application of copula-based models in travel demand modeling include but not limited to (*14-17*). Rich set of introduced copula classes, including the Gaussian, Clayton, Gumbel, Frank, and Joe copulas, allow researchers to capture the most appropriate dependency structure between random variables (*14*).

The proposed model offers a powerful tool to particularly assess the impact of TDM policies on various population segments and presents substantial benefits for planning agencies in practice. Specifically, the proposed modeling approach is capable to form the core of the trip planning module in the ADAPTS activity-based travel demand model (*18-21*) and so, it can replace the current sub-models with independent mode and departure time choices. Furthermore, since trip purpose has an indubitable effect on travel mode and departure time decisions, we focus on the home-based non-mandatory trips in

this study. In fact, several studies have focused on mandatory work trips because of direct impact on peak period congestion.

For instance, Habib et al. (*22*) estimated a joint model of mode choice and trip timing for commuting trips in the Greater Toronto Area (GTA). Habib (*7*) later added activity duration to the joint framework by incorporating it as an endogenous variable and concluded that the results are more accurate when the work duration is considered in the model. Paleti et al. (*23*) presented a joint model of mode choice and time-of-day for work trips using both observed and stated preference data. They argued that using only observed data would result in less model sensitivity to travel demand policies. On the other hand, Kumar and Levinson (*24*) stated that advances in telecommunication should enable more work-at-home, which result in more flexibility for non-mandatory activities during the day. In addition, while the possibility of shifting time of trip and/or changing the travel mode in response to TDM plans is much higher in non-mandatory trips (*1, 2*), limited attention has been devoted to modeling of their mode and timing decisions.

The performance of the proposed cluster-based joint model is compared with three other models: (i) an aggregate joint model on all non-clustered observations to assess the potential benefits of population clustering, (ii) a cluster-based univariate mode choice model, and (iii) a cluster-based univariate departure time model to analyze the potential advantages of joint modeling approach. The goodness-of-fit measures and prediction accuracy results in all these cases demonstrate that the proposed cluster-based joint model significantly outperforms the other estimated models. Furthermore, the variations in the estimated parameters for the different clusters indicate significant behavioral differences across clusters. Among the copula functions explored in this study (i.e., Frank, Gumbel, Clayton, and Joe *(11)*) to estimate the dependent structure of decisions, the model with Frank copulas provides the best statistical fit.

The rest of this study is structured as follows. Section 2 describes the data preparation process. This is followed by description of modeling approach. In section 4, detailed model estimation results and behavioral implications are presented. The study concludes with a summary of the major findings and recommendations for future studies.

2. DATA PREPARATION

The main data source used in this study is the Travel Tracker Survey that was conducted between January 2007 and February 2008 by the Chicago Metropolitan Agency for Planning (CMAP). The dataset is collected by surveying approximately 10,500 households, in which each member was asked to fill a detailed travel diary for one or two assigned dates. The dataset contains more than 210,000 trip observations and their detailed information including trip purpose, origin and destination, mode, and departure time along with demographics of the travelers.

The survey dataset was first analyzed, and invalid records were eliminated. Moreover, home-based non-mandatory trips were formed and linked to the socio-demographic information of travelers. Only home-based trips, with trip origin at home, are examined in this study, since modes of non-home-based trips are highly associated with the mode of their respective home-based trip within their tour. As for trip purposes, the non-mandatory trips are considered in this study and categorized as the following groups:

- Personal trips, such as religious, healthcare, and civic activities.
- Discretionary trips, such as dining out, visiting friends, and entertainments.
- Shopping trips, such as grocery shopping.

There are in total about 31,000 non-mandatory trips in the CMAP dataset. A sample of 11,000 trips of 1,642 individuals that follow almost the same age distribution as in the whole dataset is selected for the purpose of this study.

This study employs the choice set generation methodology proposed by Javanmardi et al. (*25*) to generate personalized attributes (e.g. travel time, travel cost) of non-chosen alternatives that are available for travelers at their departure times. They developed an application software to query the attributes of trips from Google Maps API and RTA trip planner in the same day of week and at the exact time of day that the original trips are made. This information includes exact point-to-point travel times, available modes, disaggregate access and egress distances, the nearest available transit stations/stops, and transfer information. Relevant resource constraints such as vehicle availability in the household or transit availability at the time of the observed trips are also investigated to find out the available modes for the traveler. In addition, multiple variables such as population density, transit density, and road density of TAZs are also generated to be used as proxies for land use and built environment factors. For more information regarding the data collection procedure, the reader is referred to Javanmardi et al. (*25*). Summary statistics of the key variables used in this study are presented in Table 1.

3. MODEL SPECIFICATION

This study aims to develop a joint model of travel mode and departure time choices, which takes into account the heterogeneity of travelers' decision-making behavior. To this end, individuals are assigned to clusters using a cluster analysis technique, in a way that members of each cluster are relatively homogenous in terms of their lifestyle specifications. A joint copula-based discrete-continuous model is then estimated to investigate the interrelated decision mechanism of these two travel dimensions within each cluster. In this section, first, the preparatory modeling steps including Principal Component Analysis (PCA) and cluster analysis are presented. Following that, the joint modeling approach applied in this study is detailed.

3.1. Principal Component Analysis

A descriptive analysis of the data reveals that several variables introduced in the last section are highly correlated with each other. For example, income level and education or number of vehicles and number of workers in a household are highly correlated. This possible multicollinearity between variables might cause misapprehension in identifying the influence of explanatory variables on travel behavior. To cope with multicollinearity issue, this study applies the PCA technique, which reformulates a set of observed variables into a new set (usually fewer in numbers) of independent variables (*9*). Indeed, PCA attempts to determine the components of the data that reduce the dimensions of variations and may be given a possible meaning (*26*).

A total of 27 explanatory variables of individuals' and households' demographics and built-environment characteristics are chosen for the PCA. Models with 5 to 15 components are tested with different combinations of explanatory variables, and finally 8 components are selected that can explain 76% of the variance in the dataset. The transformation is conducted using the Varimax method, a type of orthogonal rotation, with Kaiser Normalization to derive uncorrelated factors. All of the eigenvalues for the calculated factors are greater than one. The factor loadings of each of the explanatory variables onto each of the factors are summarized in Table 2.

TABLE 1 Description of Variables and Summary Statistics

Variable	Description	Mean	St. dev.
Population_density	Population density of home TAZ (population/area)	0.01	0.01
Road_density	Road density of home TAZ (roads lengths/area)	0.16	0.11
Housing_density	Housing density of home TAZ (No. of houses/area)	0.005	0.01
Transit_density	Transit density of home TAZ (No. of stops/area) ($\times 10^6$)	1.68	3.58
CBD	1: if traveler lives in CBD; 0: otherwise	0.12	0.32
Walk_TT	Travel time for walk mode (in hours)	2.45	2.93
Bike_TT	Travel time for bike mode (in hours)	0.85	0.65
Drive_TT	Travel time for auto drive mode (in hours)	0.27	0.38
Transit_TT	Travel time for transit mode (in hours)	0.44	0.32
Drive_cost	Travel cost for auto drive mode ($)	1.16	1.39
Transit_cost	Travel cost for transit mode ($)	1.79	1.60
Walk_accessible	1: if walking distance to the destination is less than 0.25 mile	0.09	0.29
Transit_egress	Egress distance to destination for transit mode (km)	0.81	1.12
Transit_access	Access distance from origin for transit mode (km)	1.26	2.09
Activity_dur	Duration of activity at trip destination (in minutes)	210.11	216.93
Weekend	1: if the trip is made in weekend; 0: otherwise	0.11	0.31
Low_income	1: if household income is less than $50,000; 0: otherwise	0.39	0.49
Med_income	1: if income is between $50,000 and $100,000; 0: otherwise	0.46	0.50
High_income	1: if income is more than $100,000; 0: otherwise	0.15	0.36
HH_bikes	Number of bikes in the household	1.38	1.67
HH_license	Number of licensed drivers in the household	1.88	0.86
HH_size	Household size	2.69	1.36
HH_worker	Number of workers in the household	1.53	0.92
HH_student	Number of students in the household	0.76	1.08
HH_vehicle	Number of vehicles in the household	1.82	1.04
Part_work	1: if traveler works part time; 0: otherwise	0.14	0.34
Full_work	1: if traveler works full time; 0: otherwise	0.53	0.51
Age –16	1: if traveler's age is less than 16; 0: otherwise	0.06	0.22
Age_17–30	1: if traveler's age is between 17 and 30; 0: otherwise	0.12	0.29
Age_31–50	1: if traveler's age is between 31 and 50; 0: otherwise	0.38	0.46
Age_51–65	1: if traveler's age is between 51 and 65; 0: otherwise	0.29	0.44
Age +65	1: if traveler's age is greater than 65; 0: otherwise	0.15	0.36
Age –20	1: if traveler's age is less than 20; 0: otherwise	0.08	0.28
Age _40–65	1: if traveler's age is between 40 and 65; 0: otherwise	0.67	0.47
White_ethnicity	1: if traveler is of white origin; 0: otherwise	0.41	0.49
Black_ethnicity	1: if traveler is of black origin; 0: otherwise	0.36	0.48
Hisp/other_ethnicity	1: if traveler is of Hispanic or other origins; 0: otherwise	0.23	0.42
No_high_degree	1: if traveler is not a high school graduate; 0: otherwise	0.09	0.29
High_degree	1: if traveler has high school degree; 0: otherwise	0.16	0.37
College_degree	1: if traveler has college degree; 0: otherwise	0.21	0.41
Bachelor_degree	1: if traveler has a bachelor degree; 0: otherwise	0.28	0.44
Grad_degree	1: if traveler has a graduate degree; 0: otherwise	0.26	0.44

TABLE 2 Principal Component Analysis Results

Variable	HH General Info	Highly Dense Area	Lower Income & Education	Higher Income & Education	Middle-age Family	Youngers	Black and Hispanic	Seniors
HH_vehicle	**0.704**	-0.320	-0.160	-0.065	-0.095	0.212	0.062	0.027
HH_size	**0.806**	0.307	0.119	-0.144	0.132	0.053	0.189	-0.089
HH_worker	**0.661**	0.053	-0.346	-0.092	-0.076	0.237	0.235	-0.012
HH_student	**0.655**	0.424	0.178	-0.116	0.193	-0.069	-0.090	-0.104
HH_license	**0.729**	-0.160	-0.242	-0.132	-0.080	0.320	-0.398	-0.003
HH_bikes	**0.548**	0.240	-0.029	-0.002	0.090	-0.243	0.024	-0.083
CBD	-0.141	**0.184**	-0.173	0.105	-0.046	0.047	-0.093	0.163
Population_density	-0.357	**0.763**	-0.318	-0.166	-0.146	0.124	-0.088	-0.010
Housing_density	-0.410	**0.748**	-0.324	-0.145	-0.143	0.108	-0.002	-0.001
Transit_density	-0.397	**0.617**	-0.218	-0.001	-0.053	0.026	0.184	-0.014
Road_density	-0.077	**0.438**	-0.146	-0.196	-0.128	0.154	-0.005	-0.069
Low_income	-0.453	0.135	**0.510**	-0.032	0.389	0.270	0.236	-0.083
No_high_degree	0.373	0.474	**0.631**	-0.204	-0.220	-0.058	0.135	0.099
High_degree	-0.157	-0.200	**0.417**	-0.216	0.210	0.182	0.284	-0.035
College_degree	-0.118	-0.182	**0.496**	-0.323	-0.019	0.066	0.032	-0.038
White_ethnicity	0.434	-0.104	**-0.585**	0.086	0.503	0.377	-0.056	-0.045
Med_income	0.119	-0.211	-0.025	**0.596**	-0.565	-0.433	0.083	-0.001
High_income	0.363	0.091	-0.406	**0.649**	0.215	0.191	-0.024	0.080
Grad_degree	-0.067	-0.022	-0.444	**0.459**	-0.127	-0.169	-0.289	-0.443
Age –16	0.362	0.438	-0.237	0.262	**0.594**	-0.191	0.298	0.087
Age_31–50	0.135	0.097	-0.437	-0.239	**0.609**	-0.447	-0.159	-0.181
Age_51– 65	-0.176	-0.306	-0.233	0.249	**0.556**	0.187	-0.248	-0.008
Age_17–30	0.142	0.107	0.034	-0.287	-0.003	**0.585**	0.328	0.120
Bachelor_degree	-0.025	-0.061	-0.329	-0.151	0.162	**0.838**	-0.021	0.202
Black_ethnicity	-0.043	0.149	0.454	-0.268	0.044	-0.029	**0.736**	-0.082
Hisp/other_ethnicity	0.049	0.094	0.289	-0.373	0.075	-0.045	**0.564**	-0.032
Age +65	-0.131	-0.280	0.254	0.006	0.126	0.020	0.035	**0.425**

NOTE: Dominant variables in boldface.

These 8 factors provide an initial understanding of the interdependencies between each of the variables and can be identified as follows:

1. Household general information. This factor represents household's six dominant variables including number of vehicles, workers, licensed drivers, bikes, and members of the households. These variables are highly correlated in the dataset.
2. Highly dense area. This factor represents dense urban areas with high population, housing, road and transit densities. It defines dimensions relating to land use.
3. Lower income and education. This factor consists of variables representing individuals with low income (<50,000$) who are mostly uneducated or have high school diploma or college degree.
4. Higher income and education. This component is affected by three dominant variables that are representing individuals with medium or high income (>50,000$) that mostly have graduate degree.

5. Middle-aged family. This factor refers to families with young to middle-aged (31-65 years old) parents having children.
6. Youngers. This factor represents young singles or couples with bachelor's degree.
7. Black and Hispanic. The variables indicating the minorities are dominant parameters in this factor.
8. Seniors. This component defines the characteristics of the senior persons (over 65 years old).

3.2. Cluster Analysis

Many studies (e.g., *27, 28*) show that travelers' life style, attitudes, and perceptions significantly affect their travel behavior, and subsequently heterogeneous travelers may response differently towards transport policies. Cluster analysis is one of the most widely used approaches to deal with heterogeneity of travelers in terms of decision-making criteria toward various choices (*29*). In this work, distance-based clustering method is used that classify observations into relatively homogenous collections by minimizing the variance across variables of interest within clusters and maximizing the variance between clusters (*9*).

Numerous clustering techniques have been proposed in the literature, among which *K*-means, hierarchical, and two-step clustering techniques are widely used in travel behavior studies (*28, 30, 31*). This study applies the two-step method because of accuracy of its reported results. Further, it is generally considered to be more efficient when dealing with large datasets that contain both continuous and categorical variables (*32*). In this method, the data is firstly divided into sub-clusters based on log-likelihood distance. It checks all the records one by one and decides whether the current observation should be merged with the previously formed clusters or assigned to a new cluster based on the distance criterion. In the second step, the resulting sub-clusters are further categorized into the desired number of clusters by comparing their distance measures with a specified threshold. If the decrease in log-likelihood as a result of merging the two clusters is larger than the threshold, the two sub-clusters can be merged. The distance between two clusters can be calculated as follows (*32*):

$$d(i,j) = \eta_i + \eta_j - \eta_{<i,j>} \tag{1}$$

where:

$$\eta_v = -N\left(\sum_{k=1}^{K_{CO}} \frac{1}{2}\log(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2) + \sum_{k=1}^{K_{CA}} \hat{E}_{vk}\right) \tag{2}$$

$$\hat{E}_{vk} = -\sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v}\log\frac{N_{vkl}}{N_v} \tag{3}$$

where K_{CO} and K_{CA} are total number of continuous and categorical variables, respectively, L_k represents number of categories in kth categorical variables, N_v is number of observations in cluster v, $\hat{\sigma}_k^2$ and $\hat{\sigma}_{vk}^2$ are the estimated variance of kth continuous variable in all data and in cluster v, respectively. N_{vkl} is number of observations in cluster v whose kth categorical variable takes the lth category, and $<i, j>$ is the index that represents the cluster formed by combining clusters i and j.

Using the factors derived from the PCA that reflect the socioeconomics and land use characteristics, the two-step clustering method is applied with 10% noise allowance to categorize the records into homogenous clusters. This optimal number of clusters is determined using Bayesian information criterion (*32*). At the end, the records are assigned to six clusters, each containing between 7.9% and 22.5% of the dataset. These results can be informative in defining various lifestyles that are assumed to influence travelers' behavior. Table 3 presents a brief summary statistic and centroids of clusters, which represent the following lifestyles:

TABLE 3 Cluster Statistics and Centers

Cluster No.	Cluster Name	Size	%	Components							
				General HH Info	Highly Dense Area	Lower Income & Edu.	Higher Income & Edu.	Middle-aged Family	Youngers	Black & Hispanic	Seniors
1	Forever worker	1,557	14.2	0.24	1.07	2.35	-0.80	-0.66	-0.59	0.17	0.67
2	Affluent in suburb	2,365	21.5	1.43	-0.29	-0.57	1.54	1.71	-0.12	-0.67	0.36
3	Young achievers	2,187	19.9	-0.36	1.81	-0.53	0.83	-0.23	2.08	0.08	-0.44
4	Seniors	869	7.9	-0.04	0.25	0.89	-0.30	0.12	-0.34	0.21	1.75
5	Mainstream families	2,462	22.7	0.84	-0.46	1.66	-0.11	1.72	0.18	-0.41	0.25
6	Minorities	1,510	13.9	0.62	-0.32	1.04	-0.82	0.99	-0.52	1.60	0.58
Total		11,000	100								

1. *Forever worker.* This cluster represents individuals with lower levels of income and education. They usually live in dense urban areas and are mostly full-time workers.
2. *Affluent in suburbs.* This group comprises middle-aged persons with higher income. They usually live in non-dense areas like suburbs. They have bachelor or graduate degrees and mostly have kids.
3. *Young achievers.* The cluster is made up of young singles or couples mostly with high education levels that live in the dense urban areas like downtown.
4. *Seniors.* This cluster consists of senior individuals who are mostly retired or work part time with low income. They are mostly white, but some black or Hispanic are included.
5. *Mainstream families.* This group represents white, middle-aged, working-class couples. They mostly have low income and have young kids. The majority live in rural areas.
6. *Minorities.* This cluster represents black, Hispanic, or people with other races. The majority do not have graduate degrees. They mostly live in upper middle-aged families and have lower income levels.

3.3. Joint Model of Travel Mode and Departure Time

As indicated earlier, a copula-based approach is applied in this study to jointly estimate trip departure time and mode choice decisions within each homogenous cluster of travelers, identified in the last sub-section. As the first component of this joint structure, a multinomial logit model is applied to predict the mode choice decision. The utility function of the choices can be written as (*33*):

$$U_{ai} = V_{ai} + \varepsilon_{ai} = \beta_a x_{ai} + \varepsilon_{ai} \quad (4)$$

where U_{ia} is person-specific utility of mode a for individual i, V_{ai} is the systematic utility, which is a function of a set of explanatory variables (x_{ai}) and corresponding parameters of weighting factors (β_a), and the random variable ε_{ai} is the error term of the utility corresponding to unobserved factors, which is assumed to have a standard type-1 extreme value distribution. Based on this assumption for the error terms, the closed form probability for selecting alternative a by person i would be expressed as (*33*):

$$P_{ai} = \frac{\exp(\beta_a x_{ai})}{\sum_k \exp(\beta_k x_{ki})} \quad (5)$$

Departure time as the second component of this model is treated as a continuous variable in this study and is modeled with a log-linear regression as:

$$ln(t_{ai}) = \alpha_a Z_{ai} + \zeta_{ai} \quad (6)$$

where $ln(t_{ai})$ represents the natural logarithm of trip timing for person i and alternative a, only if choice a is selected as the trip mode, α is the vector of estimable parameters, Z is the vector of explanatory variables, and ζ is the error term corresponding to unobserved factors.

The linkage between mode choice and trip timing decisions depends on the type and the extent of the dependency between the stochastic terms ε_{ai} and ζ_{ai}. This study applies the copula approach to capture this dependency between these two decisions. The copula presents the joint probability distribution of random variables with pre-defined marginal distributions as follows (*34*):

$$F_{\varepsilon_{ai},\zeta_{ai}}(X_1, X_2) = C_\theta\left(u_1 = F_{\varepsilon_{ai}}(X_1), u_2 = F_{\zeta_{ai}}(X_2)\right) \quad (7)$$

where $C_\theta(.,.)$ is the relevant copula function, $F_{\varepsilon_{ai},\zeta_{ai}}(.,.)$ is the multivariate joint distribution, $F_{\varepsilon_{ai}}(.)$ and $F_{\zeta_{ai}}(.)$ are marginal distributions, and θ is the dependence parameter.

In this study, four different copula functions of Frank, Gumbel, Clayton, and Joe *(11)* were explored to estimate the dependent structure, and the model with Frank copulas provided the best statistical fit. Although we are unable to provide detailed discussion of these models here due to space constraint, a comprehensive discussion of these copula structures can be found in Bhat and Eluru (*11*). Therefore, the model specifications and estimation results of only Frank copula are presented here. The copula function for Frank copula with θ as the copula parameter is as follows (*11*):

$$C_\theta(u_1, u_2) = -\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1}\right) \quad (8)$$

This joint distribution is then used to form the likelihood function. Following Spissu et al. (*35*), the likelihood function would be formulated as follows:

$$L = \prod_{i=1}^{N}\left[\left\{\prod_{a=1}^{A}\frac{1}{\sigma_{\zeta_{ai}}} \times \frac{\partial C_{\theta a}(u_{i1}^a, u_{i2}^a)}{\partial u_{i2}^a} f_{\zeta_{ai}}\left(\frac{ln(t_{ai}) - \alpha_a Z_{ai}}{\sigma_{\zeta_{ai}}}\right)\right\}^{R_{ai}}\right] \quad (9)$$

where R_{ai} is the binary variable representing whether mode a is selected by person i, $f_{\zeta_{ai}}$ is the probability density function of ζ, $\sigma_{\zeta_{ai}}$ is the scale parameter of ζ, $C_{\theta a}$ is the copula corresponding to $F_{\varepsilon_{ai},\zeta_{ai}}(u_{i1}^a, u_{i2}^a)$ with $u_{i1}^a = F_{\varepsilon_{ai}}(\beta_a x_{ai})$ and $u_{i2}^a = F_{\zeta_{ai}}\left(\frac{ln(t_{ai}) - \alpha_a Z_{ai}}{\sigma_{\zeta_{ai}}}\right)$. This likelihood function is maximized using SAS statistical software *(36)* and the variables' coefficients in discrete and continuous components as well as the copula parameters are estimated. It should be noted that after the clustering procedure, the members of each cluster are divided into training set (for model estimation) and test set (for model validation) with 80% and 20% of the observations, respectively.

4. ESTIMATION RESULTS AND DISCUSSION

Following the PCA and clustering steps, six copula-based joint models are estimated to determine the significant factors in the joint trip mode and departure time decisions of the travelers within each homogenous cluster. All possible variables and variable interactions were tested and the statistically significant coefficients (at confidence levels 90%, 95%, and 99%) are presented in Table 4. In addition, Table 5 presents the marginal effects of variables on mode choice in each cluster.

TABLE 4 Model Estimation Results for Each of the Clusters

Variables	Forever worker		Affluent in suburbs		Young achievers		Seniors		Mainstream families		Minorities		All Clusters	
	Param.	t-Stat	Param.	t-Stat	Param.	t-Stat	Param.	t-Stat	Param.	t-Stat	Param.	t-Stat	Param.	t-Stat
Mode Choice Model														
Walk_TT	-0.78***	-4.63	–	–	-2.29***	-7.08	-1.54***	-9.24	-0.52***	-3.64	-0.33***	-4.95	-1.65***	-16.76
Walk_accessible	2.1***	4.31	1.75***	8.55	0.89*	1.73	2.63***	13.45	1.89***	8.11	1.59***	7.84	2.05***	17.17
Walk_age –20	1.53***	4.43	–	–	0.83**	2.50	–	–	1.76***	7.76	1.39***	8.62	0.73***	6.28
Walk_Low_income	-0.43*	-1.82	–	–	–	–	–	–	–	–	–	–	–	–
Bike_constant	-1.84***	-3.72	-3.2***	-6.42	-3.56***	-8.74	-0.67***	-2.97	-1.84***	-6.14	-1.4***	-5.86	-2.11***	-13.89
Bike_HH_bikes	0.42***	7.44	0.43***	6.81	0.41***	4.37	0.23***	4.89	0.44***	8.61	0.41***	10.40	0.47***	21.60
Bike_TT	-1.16***	-5.39	–	–	-0.83**	-2.19	–	–	-1.9***	-5.78	-1.23***	-6.60	-2.11***	-14.10
Bike_ age _40 – 65	–	–	-0.48*	-1.91	–	–	-1.28***	-4.80	-0.96***	-3.00	-0.53**	-2.33	-0.56***	-5.20
Drive_constant	-0.9**	-1.97	-1.85***	-4.62	-2.05***	-4.32	0.58***	3.94	-0.71***	-4.01	-0.46***	-2.59	-0.09	-0.74
Drive_ HH_vehicle	1.24***	15.48	1.15***	17.00	0.72*	1.83	0.91***	12.66	1.09***	13.63	1.15***	17.47	1.16***	33.70
Drive_TT	-0.11**	-2.07	-0.38*	-1.78	-0.18***	-4.02	-0.23**	-2.32	-0.62***	-2.78	-0.29***	-4.35	-0.37**	-2.45
Drive_ age –20	-0.69***	-4.44	-0.06*	-1.69	–	–	–	–	-0.89***	-2.90	-0.64**	-2.02	-1.62***	-14.56
Drive_cost	-0.12***	-4.14	-0.05**	-2.16	-0.09**	-2.49	-0.11***	-3.24	-0.11***	-3.42	-0.15***	-5.84	-0.16***	-11.10
Transit_constant	0.44*	1.77	-2.5***	-3.47	-1.88***	-5.32	–	–	–	–	0.47**	2.36	-0.23***	-27.15
Transit_TT	-0.26*	-1.93	-0.05***	-3.45	-0.09***	-2.68	-0.06***	-2.71	-0.42*	-1.77	-0.35**	-2.63	-0.61***	-4.84
Transit_cost	-0.16***	-2.93	–	–	-0.05**	-2.23	-0.18***	-4.08	-0.11**	-2.14	-0.12***	-2.81	-0.66***	-10.28
Transit_egress	-0.69**	-2.53	-0.21***	-3.68	-0.7***	-2.95	-0.64***	-3.11	-0.37**	-2.73	-0.45**	-2.04	-0.06**	-2.54
Transit_access	–	–	-0.52***	-3.46	-1.38***	-4.00	-1.82***	-9.79	–	–	–	–	-1.18***	-17.01
Transit_ low_income	0.06**	2.11	–	–	–	–	–	–	–	–	0.09**	1.98	0.07***	1.72
Departure Time Model														
Walk_constant	6.91***	64.15	6.79***	45.91	6.98***	71.30	6.56***	108.00	6.67***	101.95	6.58***	197.72	6.72***	229.12
Walk_age –20	0.09*	1.68	-0.27***	-2.90	–	–	–	–	-0.11**	-2.48	-0.08*	-1.71	–	–
Walk_TT	-0.29***	-4.56	-0.18***	-3.74	-0.35**	-2.37	-0.28*	-1.69	-0.37***	-3.82	-0.17**	-2.56	-0.28***	-3.38
Walk_ activity_dur	-0.12***	-5.48	–	–	-0.09**	-2.15	-0.08**	-6.39	–	–	-0.11***	-4.69	-0.06***	-24.87
Bike_constant	6.88***	73.23	6.81***	44.11	6.95***	111.45	6.61***	46.91	6.71***	60.04	6.54***	93.00	6.53***	139.45
Bike_TT	-0.34***	-7.53	-0.39***	-4.78	-0.36**	-1.99	-0.46***	-2.68	-0.26*	-1.75	-0.24**	-2.06	-0.31***	-4.37
Bike_activity_dur	-0.10**	-1.99	–	–	-0.13*	-1.83	–	–	-0.12***	-3.81	-0.10**	-2.31	-0.06***	-11.44
Bike_ age –20	0.32**	2.42	-0.39***	-5.95	–	–	–	–	-0.28*	-1.77	-0.16***	-3.69	–	–
Bike_ part_work	-0.06*	-1.88	–	–	-0.06**	-2.00	–	–	–	–	-0.13***	-2.58	–	–
Drive_constant	6.93***	81.24	6.89***	91.43	7.02***	57.66	6.71***	61.52	6.75***	113.17	6.63***	116.37	6.25***	409.74
Drive_ HH_vehicle	0.25***	3.91	0.23**	2.13	–	–	0.17*	1.82	0.16***	3.59	0.14**	2.43	0.15***	24.73
Drive_TT	-0.87***	-6.35	-0.81***	-2.79	-0.65**	-2.23	-0.62*	-1.71	-0.76*	-1.80	-0.59***	-3.88	-0.03**	-2.31
Drive_ activity_dur	-0.06***	-5.04	-0.07*	-1.73	-0.01**	-2.22	-0.08**	-2.37	-0.04***	-4.51	-0.07*	-1.86	-0.06***	-45.16
Drive_ full_work	0.29***	4.61	0.37***	6.02	0.17***	4.54	–	–	0.27**	2.09	0.13***	5.65	–	–
Drive_ road_density	–	–	–	–	-1.64**	-2.29	–	–	-0.98*	-1.79	–	–	-0.10***	-2.87
Transit_constant	6.83***	120.91	6.75***	42.54	6.88***	55.19	6.84***	120.47	6.75***	148.53	6.82***	258.25	6.36***	336.76
Transit_ HH_size	-0.26***	9.25	–	–	–	–	–	–	-0.07***	-3.49	-0.16***	-10.48	0.03***	4.64
Transit_ age +65	–	–	–	–	–	–	-0.67***	-7.19	-0.49**	-2.02	–	–	–	–
Transit_TT	-0.61***	3.06	-0.49***	-4.67	-0.54***	-8.57	-0.53***	-8.27	-0.51***	-5.38	-0.62***	-6.09	–	–
Transit_ activity_dur	-0.15***	4.31	-0.11***	-9.57	-0.21***	-4.25	-0.18**	-2.17	-0.14***	-6.48	-0.13**	-2.49	-0.03***	-13.74
Transit_ Weekend	-0.17*	1.68	–	–	-0.26**	-2.38	–	–	-0.24**	-2.68	-0.31***	-4.82	–	–
Copula Parameters														
θ_{Walk}	-6.39***	-4.87	-5.55***	-6.76	-3.26**	-2.27	-0.48***	-3.15	-0.51***	-2.89	-0.67***	-5.38	-0.48***	-10.25
θ_{Bike}	-6.01***	-2.59	-7.62***	-5.37	-2.56***	-3.79	-0.49***	-3.49	-0.54***	-3.76	-0.69***	-5.14	-0.47***	-12.44
θ_{Drive}	-6.18***	-16.80	-9.89***	-4.14	-7.24***	-4.44	-0.42**	-2.26	-0.83***	-6.36	-1***	-7.56	-0.54***	-14.66
$\theta_{Transit}$	-8.11***	-9.34	-13.9***	-20.57	-4.65***	-5.56	-0.55***	-8.16	-0.63***	-9.46	-0.77***	-8.90	-0.64***	-38.39
Scale Parameters														
σ_{Walk}	1.03***	8.13	4.49***	15.34	1.02***	4.78	1.15***	11.43	1.00***	9.62	1.00***	11.70	1.00***	27.39
σ_{Bike}	2.94***	14.59	3.74***	14.31	1.00***	2.84	1.34***	6.47	1.19***	5.49	1.14***	7.02	1.18***	15.78
σ_{Drive}	1.78***	49.99	1.82***	54.35	1.00***	13.02	1.06***	25.20	1.00***	19.51	1.00***	25.15	1.00***	59.74
$\sigma_{Transit}$	1.00***	12.61	1.68***	14.17	1.00***	11.43	1.00***	16.34	1.00***	14.01	1.00***	16.53	1.00***	38.46
Number of observations	1557		2365		2187		869		1550		2472		11000	
Log-likelihood	-2683.71		-4183.31		-3976.58		-1928.24		-2590.64		-5332.86		-18092.79	

NOTE: *Significant at 90% , **significant at 95% , ***significant at 99%

TABLE 5 Marginal Effects for Model Variables in all Clusters

Variables	Forever worker				Affluent in suburbs				Young achievers			
	Walk	Bike	Drive	Transit	Walk	Bike	Drive	Transit	Walk	Bike	Drive	Transit
Walk_TT	-0.050	0.005	0.028	0.017	–	–	–	–	-0.130	0.010	0.058	0.062
Walk_accessible	0.251	-0.020	-0.153	-0.078	0.209	-0.011	-0.146	-0.052	0.027	-0.002	-0.013	-0.012
Walk_age –20	0.157	-0.013	-0.095	-0.049	–	–	–	–	0.083	-0.006	-0.040	-0.038
Walk_Low_income	-0.032	0.003	0.018	0.011	–	–	–	–	–	–	–	–
Bike_HH_bikes	-0.003	0.020	-0.011	-0.006	-0.006	0.013	-0.005	-0.002	-0.003	0.021	-0.013	-0.005
Bike_TT	0.005	-0.029	0.016	0.008	–	–	–	–	0.004	-0.025	0.015	0.006
Bike_ age _40 – 65	–	–	–	–	0.005	-0.012	0.005	0.001	–	–	–	–
Drive_ HH_vehicle	-0.052	-0.023	0.179	-0.104	-0.162	-0.010	0.219	-0.047	-0.031	-0.017	0.110	-0.062
Drive_TT	0.005	0.002	-0.018	0.011	0.053	0.004	-0.071	0.014	0.008	0.005	-0.027	0.014
Drive_ age –20	0.031	0.016	-0.119	0.072	0.008	0.001	-0.011	0.002	–	–	–	–
Drive_cost	0.005	0.003	-0.020	0.012	0.007	0.001	-0.010	0.002	0.004	0.002	-0.014	0.008
Transit_TT	0.002	0.002	0.014	-0.018	0.001	0.00	0.001	-0.002	0.000	0.001	0.002	-0.003
Transit_cost	0.001	0.001	0.009	-0.011	–	–	–	–	0.000	0.001	0.001	-0.002
Transit_egress	0.006	0.003	0.034	-0.043	0.001	0.00	0.001	-0.002	0.002	0.002	0.015	-0.019
Transit_access	–	–	–	–	0.003	0.001	0.002	-0.005	0.004	0.03	0.024	-0.031
Transit_ low_income	-0.001	-0.001	-0.003	0.005	–	–	–	–	–	–	–	–

Variables	Seniors				Mainstream families				Minorities			
	Walk	Bike	Drive	Transit	Walk	Bike	Drive	Transit	Walk	Bike	Drive	Transit
Walk_TT	-0.047	0.004	0.036	0.007	-0.048	0.003	0.029	0.016	-0.031	0.003	0.019	0.009
Walk_accessible	0.047	-0.004	-0.037	-0.006	0.281	-0.015	-0.181	-0.085	0.227	-0.017	-0.147	-0.063
Walk_age –20	–	–	–	–	0.254	-0.014	-0.162	-0.078	0.190	-0.014	-0.122	-0.054
Walk_Low_income	–	–	–	–	–	–	–	–	–	–	–	–
Bike_HH_bikes	-0.001	0.013	-0.011	-0.001	-0.003	0.013	-0.007	-0.003	-0.004	0.018	-0.010	-0.004
Bike_TT	–	–	–	–	0.006	-0.024	0.012	0.006	0.006	-0.030	0.0157	0.007
Bike_ age _40 – 65	0.004	-0.062	0.052	0.006	0.009	-0.038	0.020	0.009	0.009	-0.042	0.023	0.010
Drive_ HH_vehicle	-0.027	-0.031	0.090	-0.032	-0.065	-0.013	0.171	-0.093	-0.065	-0.020	0.175	-0.090
Drive_TT	0.008	0.011	-0.029	0.010	0.040	0.008	-0.112	0.064	0.018	0.006	-0.052	0.027
Drive_ age –20	–	–	–	–	0.058	0.012	-0.162	0.092	0.060	0.020	-0.171	0.091
Drive_cost	0.004	0.005	-0.014	0.005	0.007	0.002	-0.119	0.110	0.009	0.003	-0.026	0.014
Transit_TT	0.00	0.00	0.001	-0.001	0.006	0.001	0.021	-0.028	0.005	0.002	0.017	-0.024
Transit_cost	0.00	0.00	0.002	-0.002	0.002	0.001	0.006	-0.009	0.002	0.001	0.006	-0.009
Transit_egress	0.00	0.001	0.006	-0.007	0.005	0.001	0.019	-0.025	-0.006	0.002	0.022	-0.030
Transit_access	0.001	0.002	0.011	-0.014	–	–	–	–	–	–	–	–
Transit_ low_income	–	–	–	–	–	–	–	–	0.005	0.002	0.018	-0.025

1

The estimation results indicate that various socio-demographic, land-use, and trip-related variables in both the discrete and continuous components remained statistically significant across all travelers' clusters. Furthermore, the results show that applying the clustering technique, which enables us to account for unobserved shared factors among members of each cluster, leads to variation in magnitude of influence and significance level of variables across clusters. For example, travel time has a mixed effect on choosing *Walk* as the mode of travel. The results demonstrate that generally people who face greater travel times are less inclined to walk to their destination. According to Table 5, members of cluster 3 are more sensitive to change in travel time than others; where one unit increase in travel time leads to 13% reduction in probability of choosing *Walk*. However, in the second cluster travel time has no significant effect on selecting this mode. This is possibly due to the lifestyle specifications of people in this cluster. Referring to Table 3, this cluster mainly consists of people with higher income levels who live in the non-dense areas (e.g., suburbs), where generally shopping centers or other non-mandatory trip destinations are not as accessible as in dense urban areas (e.g., CBD). Similarly, the effect of travel time on choosing *Bike* varies significantly across clusters; it is even not significant in cluster 2 (i.e., residents of non-dense areas with high income) and cluster 4, which mostly includes senior citizens. On the other hand, in utility functions of *Drive* and *Transit* modes, travel time plays a significant role in travel mode decision of travelers in all user clusters.

Moving to transit-related variables, it is found that egress distance is a main factor in choosing *Transit* as the travel mode across all clusters. The results indicate that people in all clusters are less likely to choose *Transit* in trips with longer egress distances, whereas the access distance is significant in only half of the clusters. The partial significance of this variable is possibly due to behavior of travelers who do not have access to other modes of transportation. Travel cost also has variable negative effect on *Transit* mode choice in all clusters except cluster 2 wherein it is not significant.

With respect to trip departure time models, the results highlight that travel time is one of the key contributors to trip departure time decision; that is the increase in travel time in almost all modes and users clusters results in earlier departure times except for the senior people. Activity duration is also found to significantly affect trip timing; trips with longer activities at their destinations are generally taken at earlier times. Work status of the travelers is also an important factor in estimating the trip departure time, as part time workers tend to perform non-mandatory activities sooner than the individuals with full time jobs. Turning to household socio-demographic characteristics, the result show that number of vehicles in the household and household size significantly affect the departure time; more vehicles in the household results in later departure times possibly due to reduction of shared trips in the household and subsequently more flexibility of travelers. The results also indicate that people start their trips sooner if they are destined to high-density areas.

In order to further investigate the implications of clustering approach in the modeling scheme, an aggregate joint model for the whole observations is also estimated; the results of this model are presented in the last column of Table 4. The estimated parameters show remarkable difference in magnitude of the estimated coefficients, their significance degree, and in some cases, their signs between the aggregate model and the cluster-specific models. This fact indicates the potential behavioral differences across clusters and highlights the importance of segmentation analysis in devising transportation policies, which should account for diverse responses of all travelers in the network.

To assess the prediction capability of these models, they are used to simulate selected choices in test sets. Figure 1 presents the prediction accuracy (percentage of correctly predicted alternative) of mode choice components of joint models for both cluster-based models and aggregate model. The analysis

reveals a significant improvement in prediction accuracy in the proposed clustered-based joint models across all modes (i.e., walk, bike, drive, and transit). The greatest improvement happens in the *Drive* mode with 9.08% increase in prediction accuracy and the lowest improvement is in predicting *Bike* with 1.14% increase. Lower increase in the prediction accuracy of *Bike* could be as a result of its few trip observations in the dataset, which limits the model performance in identifying bike-related travel attributes in each cluster; thus, the clustering step does not significantly improve the model prediction accuracy for this mode.

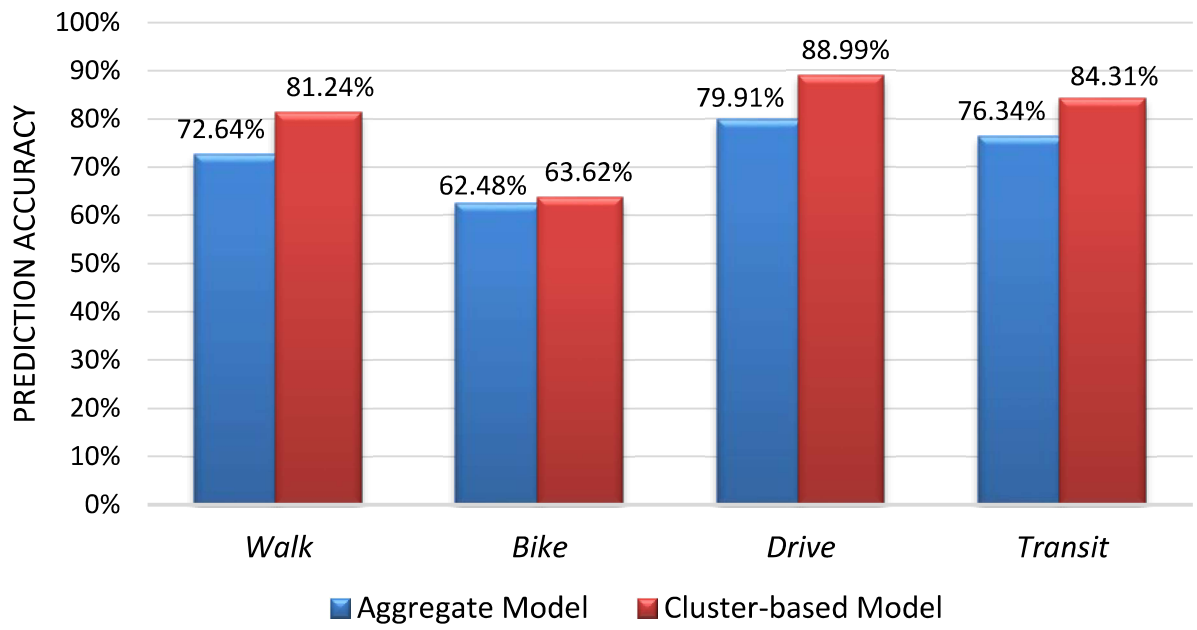

FIGURE 1 Travel mode prediction accuracy in clustered and aggregate models.

In order to analyze the potential effect of segmentation on departure time estimation, the mean absolute percentage error (MAPE) index is calculated for this component of copula models in both aggregate and clustered approaches. The MAPE index shows the average error in estimation of departure times, and is defined as (*37*):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{10}$$

where A_t is the actual value and F_t is the forecasted value for observation t. The MAPE measure for the aggregate model is 17.45% and for the cluster-based models is 10.51%, which shows a reduction of 6.94% in the average error. Figure 2 presents the predicted departure times versus the observed ones to visually support the improved prediction accuracy in cluster-based models over the aggregate one.

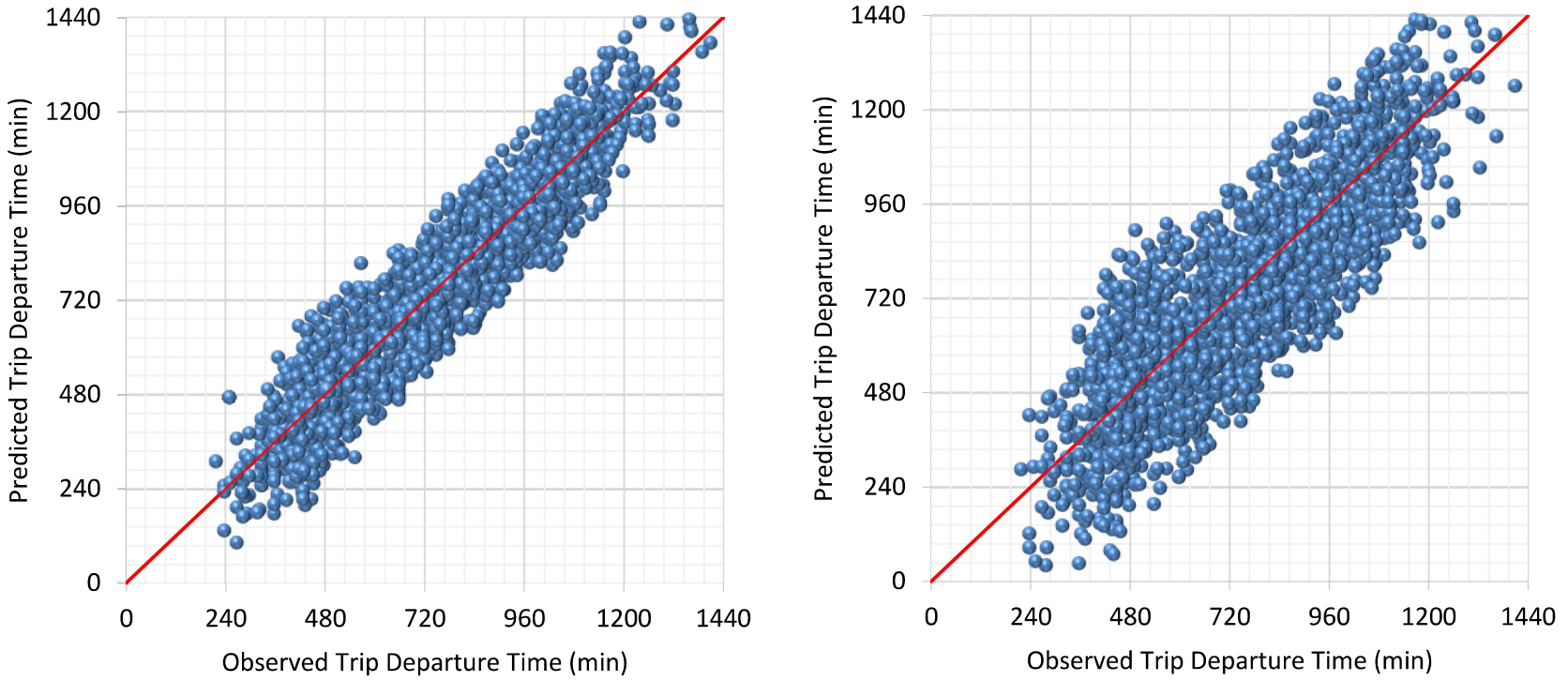

FIGURE 2 Departure time prediction accuracy for cluster-based models (left) and aggregate model (right).

Furthermore, in order to assess the benefits of considering the two decisions of mode choice and trip timing as a joint model, cluster-based univariate models are also estimated, i.e. within each users' cluster, separate MNL and log-linear regression models are estimated. The prediction accuracy of the univariate mode choice models for all clusters is presented in Table 6.

TABLE 6 Prediction Accuracy for Mode Choice Models

Cluster ID	Mode	Prediction Accuracy		Difference
		Univariate Model	Joint Model	
Cluster 1	Walk	68.31%	78.09%	9.78%
	Bike	60.77%	62.11%	1.34%
	Drive	64.38%	89.45%	25.07%
	Transit	72.44%	87.42%	14.98%
Cluster 2	Walk	73.97%	80.27%	6.30%
	Bike	62.78%	61.71%	-1.07%
	Drive	66.17%	87.16%	20.99%
	Transit	65.97%	86.89%	20.92%
Cluster 3	Walk	70.07%	84.78%	14.71%
	Bike	62.63%	66.43%	3.80%
	Drive	69.57%	89.82%	20.25%
	Transit	72.36%	81.28%	8.92%
Cluster 4	Walk	68.51%	79.26%	10.75%
	Bike	61.58%	70.84%	9.26%
	Drive	74.63%	91.90%	17.27%
	Transit	64.30%	76.31%	12.01%
Cluster 5	Walk	68.66%	82.80%	14.14%
	Bike	63.13%	61.81%	-1.32%
	Drive	65.79%	89.54%	23.75%
	Transit	65.66%	85.91%	20.25%
Cluster 6	Walk	70.75%	82.21%	11.46%
	Bike	58.26%	64.98%	6.72%
	Drive	71.64%	90.56%	18.92%
	Transit	66.28%	86.23%	19.95%

The results demonstrate significant improvement when applying the joint modeling scheme on the clusters over the univariate modeling scheme. Only in two clusters, the prediction accuracy for the *Bike* mode in univariate model is slightly better than the joint model, which might be because of the small numbers of observations for this mode (only 3% of the observations chose bike as their trip mode).

The joint modeling approach is also demonstrated to be more accurate for departure time modeling. The MAPE index of 25.27% for the clustered univariate trip departure time model shows a poor prediction accuracy in comparison with the MAPE value of 10.51% for the clustered joint model. To illustrate the goodness-of-fit of this model, predicted and observed values of departure times for the clustered univariate log-linear regression model are depicted in Figure 3. Overall, the results suggest that both segmentation technique and joint modeling approach significantly improve the prediction accuracy of mode choice and departure time models.

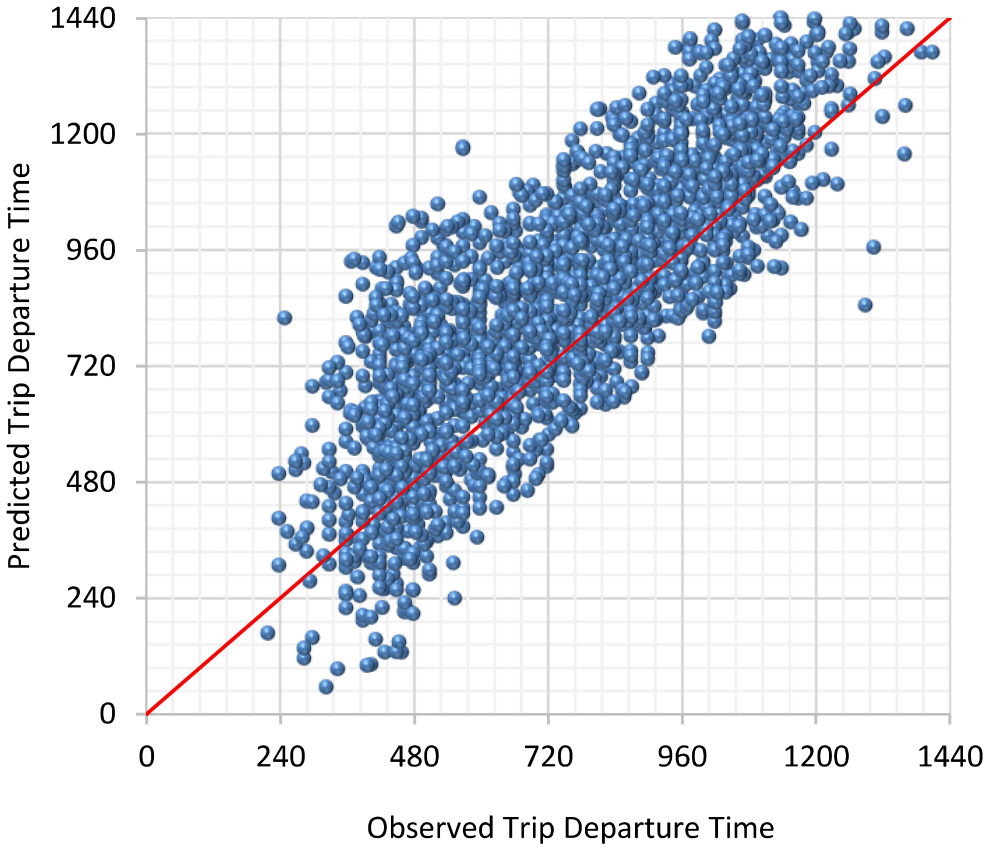

FIGURE 3 Departure time prediction accuracy for cluster-based univariate trip departure time model.

5. CONCLUSION

Recognizing the recent emergence of powerful data science techniques, this study presents an effort to analyze the travelers' behavior regarding two of the most fundamental trip-related decisions, i.e. mode choice and departure time. Due to the interrelated decision mechanism of these two travel dimensions, they should be considered jointly to allow for unrestricted correlation between their influencing factors. Moreover, because of heterogeneity in travelers' decision-making criteria toward these travel attributes, diverse travelers can respond differently to transport policies. Therefore, capturing this diversity is desirable to develop a reliable and policy sensitive model.

To achieve this goal, this study proposes a modeling approach based on cluster analysis and joint copula-based modeling technique. First, principal component analysis is applied to remove the potential multicollinearity among explanatory variables and reduce the data dimension. Following that, a two-step clustering technique is adopted to assign travelers to homogeneous clusters. Then, a copula-based joint model of discrete mode selection and continuous departure time decision is estimated within each cluster. The mode choice model is estimated by a multinomial logit model and the departure time choice is modeled by means of a log-linear regression model.

In order to investigate the potential advantages of developing different models for heterogeneous clusters, an aggregate joint model on all observations is also established. The results show that the overall clustered models outperform the aggregate one in both decision components. The estimated prediction accuracy for mode choice component is improved from 1.14% for bike mode to 9.08% for drive mode. In addition, applying the segmentation technique results in 6.94% enhancement in the prediction accuracy of trip departure time. Further, the variations in estimated parameters indicate the potential behavioral differences across clusters and highlight the importance of segmentation analysis in devising transportation policies, which should account for diverse responses of all travelers in the network.

Moreover, separate mode choice and departure time models within each cluster are also estimated to evaluate the effect of jointly considering these decisions. Comparison of prediction accuracy measures of joint models and separate ones reveals the joint approach significantly outperforms the univariate models in both mode choice and trip timing components. In addition, significant estimated copula parameters show the strong interrelationship between influencing factor of these decisions.

This study has several potentials for future research directions. First, as an extension of this work, developing a joint model that considers the correlation of other activity attributes such as activity duration with these two dimensions is desirable to better simulate travelers' trip making behavior. Also, applying other joint modeling techniques and comparing their results with the employed copula approach would be informative about their performance and coupling structures. Furthermore, future research can focus on comparison of the proposed cluster-based joint model and other methods that are able to account for heterogeneity such as random-parameters models and latent class models. Also, while some recent studies *(e.g., 38-41)* have highlighted the growing influence of social networks on travel mode choice, our study is unable to account for their role, due to lack of related information in the employed data sources. Moreover, since mandatory and maintenance activities would significantly influence the attributes of non-mandatory activities, future research can improve the model structure by incorporating new variables of mandatory and maintenance activities to capture their potential effects on non-mandatory activities.

REFERENCES

[1] Tringides, C. A., X. Ye, and R. Pendyala. Departure-Time Choice and Mode Choice for Nonwork Trips Alternative Formulations of Joint Model Systems. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1898, 2004, pp. 1–9.

[2] Bhat, C. R. Analysis of travel mode and departure time choice for urban shopping trips. *Transportation Research Part B: Methodological*, Vol. 32 (6), 1998, pp. 361–371.

[3] Bhat, C. R., and J. Y. Guo. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological*, Vol. 41 (5), 2007, pp. 506–526.

[4] Rashidi, T. H., and T. T. Koo. An analysis on travel party composition and expenditure: a discrete-continuous model. *Annals of Tourism Research*, Vol. 56, 2016, pp. 48-64.

[5] Heckman, J. Shadow prices, market wages and labor supply. *Econometrica*, Vol. 42 (4), 1974, pp. 679–694.

[6] Lee, L. F. Some approaches to the correction of selectivity bias. *Review of Economic Studies*, Vol. 49 (3), 1982, 355–372.

[7] Habib, K. M. Modelling commuting mode choice jointly with work start time and duration. *Transportation Research Part A: Policy and Practice. Vol.* 46 (1), 2012, pp. 33–47.

[8] Mohammadian, A., and Y. Zhang. Investigating transferability of national household travel survey data. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1993, 2007, pp. 67–79.

[9] Li, Z., W. Wang, C. Yang, and D. R. Ragland. 2013. Bicycle commuting market analysis using attitudinal market segmentation approach. *Transportation Research Part A: Policy and Practice*, Vol. 47, 2013, pp. 56–68.

[10] Miralinaghi, M., Lou, Y., Hsu, Y.T., Shabanpour, R., Shafahi, Y., 2016. Multiclass fuzzy user equilibrium with endogenous membership functions and risk-taking behaviors. *Journal of Advanced Transportation*. doi:10.1002/atr.1425.

[11] Bhat, C. R., and N. Eluru. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, Vol. 43 (7), 2009, pp. 749–765.

[12] Eluru, N., R. Paleti, R. M. Pendyala, and C. R. Bhat. Modeling injury severity of multiple occupants of vehicles, Copula-based multivariate approach. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2165, 2010, pp. 1–11.

[13] Born K., S. Yasmin, D. You, N. Eluru, C. R. Bhat, and R. M. Pendyala. 2014. A joint model of weekend discretionary activity participation and episode duration. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2413, 2014, pp. 34–44.

[14] Karimi, B., Z. Pourabdollahi, R. S. Anbarani, and A. K. Mohammadian. A mixed copula-based joint model of non-mandatory out-of-home activity type and activity duration. In: Proceedings 94th Annual Meeting of Transportation Research Board, Washington, D.C., 2015.

[15] Sener I. N., N. Eluru, and C. R. Bhat. On jointly analyzing the physical activity participation levels of individuals in a family unit using a multivariate copula framework. *Journal of Choice Modeling*. Vol. 3 (3), 2010, pp. 1–38.

[16] Sener, I. N., and P. R. Reeder. An integrated analysis of workers' physically active activity and active travel choice behavior. *Transportation Research Part B: Methodological*, Vol. 67, 2014, pp. 381–393.

[17] Rashidi, T. H., and A. Mohammadian. Application of a nested trivariate copula structure in a competing duration hazard-based vehicle transaction decision model. *Transportmetrica A: Transport Science*, Vol. 12 (6), 2016, pp. 1–18.

[18] Auld, J. A., and A. Mohammadian. Activity planning processes in the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) Model. *Transportation Research Part A: Policy and Practice.* Vol. 46 (8), 2012, pp. 1386–1403.

[19] Javanmardi, M., M. F. Langerudi, R. Shabanpour, and A. Mohammadian. An optimization approach to resolve activity scheduling conflicts in ADAPTS activity-based model. *Transportation,* Vol. 43 (6), 2016, pp. 1023-1039. DOI: 10.1007/s11116-016-9721-7.

[20] Fasihozaman Langerudi, M., R. Shabanpour Anbarani, M. Javanmardi, and A. K. Mohammadian. Activity Scheduling Conflict Resolution: A Reverse Pairwise Comparison of In-home and Out-of-Home Activities. *Transportation Research Record: Journal of the Transportation Research Board*, 2016, DOI: 10.3141/2566-05.

[21] Shabanpour Anbarani, R., M. Javanmardi, M. Fasihozaman Langerudi, and A. K. Mohammadian. Analyzing Impacts of Individuals' Travel Behavior on Air Pollution: Integration of a Dynamic Activity-Based Travel Demand Model with Dynamic Traffic Assignment and Emission Models. In: Proceedings 95th Annual Meeting of Transportation Research Board, Washington, D.C., 2016.

[22] Habib, K. M., N. Day, and E. J. Miller. An investigation of commuting trip timing and mode choice in Greater Toronto Area: application of a joint discrete–continuous model. *Transportation Research Part A: Policy and Practice.* Vol. 43 (7), 2009, pp. 639–653.

[23] Paleti, R., P. S. Vovsha, D. Givon, and Y. Birotker. Joint modeling of trip mode and departure time choices using revealed and stated preference data. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2429, 2014, pp. 67–78.

[24] Kumar, A., and D. M. Levinson. 1995. Temporal Variations on Allocation of Time. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1493, 1995, pp. 118–127.

[25] Javanmardi, M., M. F. Langerudi, R. Shabanpour, and K. Mohammadian. Mode choice modelling using personalized travel time and cost data. In: *International Conference of the International Association for Travel Behavior Research (IATBR)*, Windsor, UK, 2015.

[26] Stopher, P. R., and A. H. Meyburg. *Survey sampling and multivariate analysis for social scientists and engineers*. Lexinton, MA: Lexington Books, 1979.

[27] Anable, J. 'Complacent car addicts' or 'aspiring environmentalists'? Identifying travel behavior segments using attitude theory. *Transport Policy*, Vol. 12 (1), 2005, pp. 65–78.

[28] McCarthy, O. T., B. Caulfield, M. O'Mahony. Technology engagement and privacy: A cluster analysis of reported social network use among transport survey respondents. *Transportation Research Part C: Emerging Technologies*, Vol. 63, 2016, pp. 195–206.

[29] Everitt, B. S., S. Landau, and M. Leese. *Cluster Analysis* (4th edition). London, Arnold Press, 2001.

[30] Collum, K. K., and J. J. Daigle. Combining attitude theory and segmentation analysis to understand travel mode choice at a national park. *Journal of Outdoor Recreation and Tourism*, Vol 9, 2015, pp. 17–25.

[31] Pronello, C., and C. Camusso. Travellers' profiles definition using statistical multivariate analysis of attitudinal variables. *Journal of Transport Geography*. Vol. 19, 2011, 1294–1308.

[32] Chiu, T., D. Fang, J. Chen, Y. Wang, and C. Jeris. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, California, 2001.

[33] Ben-Akiva, M., and S. Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, 1985.

[34] Sklar, A. Random variables, joint distribution functions, and copulas. *Kybernetika*, Vol. 9 (6), 1973, pp. 449–460.

[35] Spissu, E., A. R. Pinjari, R. M. Pendyala, and C. R. Bhat. A copula-based multinomial discrete-continuous model of vehicle type choice and miles of travel. Transportation, Vol. 36, 2009, pp. 403–422.

[36] SAS Institute Inc. 2010. SAS/OR® 9.22 User's Guide: Mathematical Programming. Cary, NC: SAS Institute Inc.

[37] Washington, S., M. Karlaftis, and F. Mannering. *Statistical and econometric methods for transportation data analysis*, second edition. Chapman & Hall/CRC, Boca Raton, 2010.

[38] Rezende, P. H. D. R., A.M. Sadri, and S.V. Ukkusuri. Social Network Influence on Mode Choice and Carpooling during Special Events: The Case of Purdue Game Day. In: Proceedings 95th Annual Meeting of Transportation Research Board, Washington, D.C., 2016.

[39] Pike, S., Travel Mode Choice and Social and Spatial Reference Groups: Comparison of Two Formulations. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2412, 2014, pp. 75–81.

[40] Sadri, A. M., S. Lee, and S.V. Ukkusuri. Modeling Social Network Influence on Joint Trip Frequency for Regular Activity Travel Decisions. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2495, 2015, pp. 83–93.

[41] Maness, M. and C. Cirillo. An indirect latent informational conformity social influence choice model: Formulation and case study. *Transportation Research Part B: Methodological*, Vol. 93, 2016, pp. 75-101.