Hindawi
Journal of Advanced Transportation
Volume 2017, Article ID 7290248, 13 pages
https://doi.org/10.1155/2017/7290248

WILEY | Hindawi

Research Article

Prediction of Individual Social-Demographic Role Based on Travel Behavior Variability Using Long-Term GPS Data

Lei Zhu,[1] Jeffrey Gonder,[1] and Lei Lin[2]

[1] *Transportation and Hydrogen Systems Center, National Renewable Energy Laboratory (NREL), 15013 Denver West Parkway, Golden, CO 80401, USA*
[2] *Department of Civil, Structural and Environmental Engineering, University at Buffalo, Buffalo, NY 14260, USA*

Correspondence should be addressed to Lei Zhu; zhulei0717@gmail.com

Received 14 April 2017; Revised 26 June 2017; Accepted 2 July 2017; Published 16 August 2017

Academic Editor: Takahiko Kusakabe

With the development of and advances in smartphones and global positioning system (GPS) devices, travelers' long-term travel behaviors are not impossible to obtain. This study investigates the pattern of individual travel behavior and its correlation with social-demographic features. For different social-demographic groups (e.g., full-time employees and students), the individual travel behavior may have specific temporal-spatial-mobile constraints. The study first extracts the home-based tours, including Home-to-Home and Home-to-Non-Home, from long-term raw GPS data. The travel behavior pattern is then delineated by home-based tour features, such as departure time, destination location entropy, travel time, and driving time ratio. The travel behavior variability describes the variances of travelers' activity behavior features for an extended period. After that, the variability pattern of an individual's travel behavior is used for estimating the individual's social-demographic information, such as social-demographic role, by a supervised learning approach, support vector machine. In this study, a long-term (18-month) recorded GPS data set from Puget Sound Regional Council is used. The experiment's result is very promising. The sensitivity analysis shows that as the number of tours thresholds increases, the variability of most travel behavior features converges, while the prediction performance may not change for the fixed test data.

1. Introduction

An activity-travel behavior pattern analysis includes the identification of activity patterns, such as types, duration, sequence, and locations, and the recognition of travel behavior pattern regarding departure time, travel time, and travel types, such as commuting and noncommuting. It is one of the most fundamental research topics to many real-world applications, including Active Traffic and Demand Management (ATDM), Mobility-as-a-Service, and transportation demand management. The activity-travel behavior pattern is derived from either manually collected traveler activity diaries in travel surveys or passively obtained data, like global positioning system (GPS) trajectory data [1–5], geolocation data [6, 7], and transit smart card data [8, 9].

Travel demand management, such as ATDM, aims to reduce traffic demand or to redistribute the traffic demand temporally or spatially [10]. There are "hard" and "soft" strategies [11]. The "hard" strategies, also called hard policy measures, use a penalty to enforce travel behavior changes [12], including road pricing [13], toll roads [14–16], and parking pricing [17]. The "soft" measures include two categories. The first one offers traffic information to impact travelers' decisions, which does not force behavior change [18]. Implementation cases include a comparison study of passengers' travel choice behavior by altering the train timetable, proposed by Kusakabe in Japan [19, 20]; a dynamic ridesharing service, Virtual Bus in Italy [21]; a Predict-a-Trip traffic information forecast program in San Francisco [22]; and so on. The second category of "soft" measures uses incentives to influence traveler behavior and has recently attracted attention worldwide. A study in Germany showed an increase in bus use by offering prepaid bus tickets [23]. An early bird, free-ticket program, applied in Melbourne, Australia, aimed to mitigate the rail overcrowding issue and to shift the demand from peak to nonpeak hours [24]. In

2013, a 10-week pilot study was conducted by Metropia in the Los Angeles area using an incentive-based activity demand management smartphone app [10], and significant travel behavior changes, including departure time choices and route options, were observed.

For incentive strategies, the challenge is that the travel patterns and social-demographic features of the target users are not entirely understood. Some ATDM programs use incentives to influence travelers in specific groups [25, 26], like transit riders, while some apply incentives to general autodrivers directly [10]. The limited incentive resources distributed to a significant amount of general travelers may not be efficient for influencing travel behavior. To stimulate travelers to change their travel behavior efficiently and effectively, recognizing the travelers' social-demographic information, such as social-demographic roles and the associated travel pattern, scientifically dispatching incentives into the targeted individuals or specific individual groups are critical for an incentive strategy in ATDM.

However, collecting travelers' social-demographic information is not trivial. The most used method is collecting an activity diary in a traffic survey, including paper-based questionnaires and telephone interviews [27]. However, the traffic surveys usually only recruit a small number of participants for a short period (days or weeks), with shortcomings of cost, labor, and unguaranteed accuracy. Fortunately, with the prevalence of location-aware devices, such as a smartphone or GPS-enabled devices, the long-term (months or years) continuous collection of individualized trajectory data offers an unprecedented opportunity to gain insight into the traveler's daily travel pattern. Particularly, the GPS data provided by smartphone apps, such as Uber [28], Google, and Metropia [10], and the instrumented data derived by GPS devices mounted in vehicles, are among the latest sources of a new information collection mechanism. Rich information relevant to one's travel behavior is embedded in such long-term continuous collected raw GPS data. However, extracting the travel behavior patterns from raw GPS trajectory data and using them to predict an individual's social-demographic role are challenging.

Travel behavior variability describes the variance of travel behavior for an extended period, which was recognized and studied [29–31] recently. Some researchers focus on the temporal variability of travel behavior characteristics, such as daily travel time [4, 32, 33]. The spatial variability (e.g., activity locations), in which the travelers either repeat or vary their location choice over days, is also studied [32, 34]. In addition to the temporal variability and spatial variability, the mobile variability, such as driving time ratio variance and travel time variance, describes the individual's movement characteristics. The temporal-spatial-mobile variability reflects the travel characteristics with respect to time, space, and mobility. It is directly correlated with a traveler's demographic feature, especially social-demographic role (i.e., employment status), like full-time employee, part-time employee, student, retired worker, and so on [35, 36]. For example, a full-time employee is usually a daily commuter from home to work with tight departure time and destination constraints. The commuter may not have much flexibility to stop during the trip or to detour onto a different route. On the other hand, a retired worker may not be a regular commuter and has loose temporal-spatial-mobile restrictions.

This study proposes a social-demographic role prediction framework based on individuals' travel behavior variability. It first extracts travel behavior variability from a long-term GPS data set. The travel behavior variability is decomposed as three-dimensional features: temporal, spatial, and mobile. The temporal dimension represents the departure time variability, and the spatial dimension indicates the destination location variability. The fluctuations of trip travel time and driving time ratio form the mobile variability dimension. In this study, the travelers' home sites are detected from the raw GPS data. Then, the home-based tours and the travel behavior variability are produced. Next, the travel behavior variability is fed into a supervised machine learning model (support vector machine) to predict travelers' social-demographic roles. The study built upon the Puget Sound Regional Council household 2004–2006 survey data, which are provided by the National Renewable Energy Laboratory's Transportation Secure Data Center [37]. The data set includes 18 months of continuous GPS tracking over survey 450 vehicles from 275 households and the individual traveler demographic information from the travel survey. This complete data set is used not only to extract the travel behavior variability pattern of the survey respondents from their extended period continuous GPS data but more importantly to cross reference with the traditional house survey data and build machine learning models for social role prediction. Other social-demographic variables, such as income, age, and gender, are also tested to understand the general performance of the proposed social-demographic prediction model. Additionally, this study conducts a sensitivity analysis, which investigates the impact of the data collection criterion (i.e., number of tours) on tour variability and social-demographic character prediction. The major features and contributions of this research are summarized below:

(i) This study proposes an individual social-demographic role prediction model based on travel behavior variability. The travel behavior variability and its correlation to the social-demographic role are explored.

(ii) A sensitivity analysis of sampling threshold for a long-term data set reveals how the travel behavior variability and social-demographic role prediction change by different data sampling thresholds.

This research is expected to provide a practical process framework to fully take advantage of available emerging data (i.e., continuous GPS tracked data) and integrate them into the existing modeling or behavior-related research and applications. These are elaborated in the following sections. The details of travel behavior variability extraction and the social role prediction method are introduced in Methodology. Case Study and Discussion describe the experimental details and the experimental results on the testing data set. It also reveals the result of the sensitivity study of the impact of

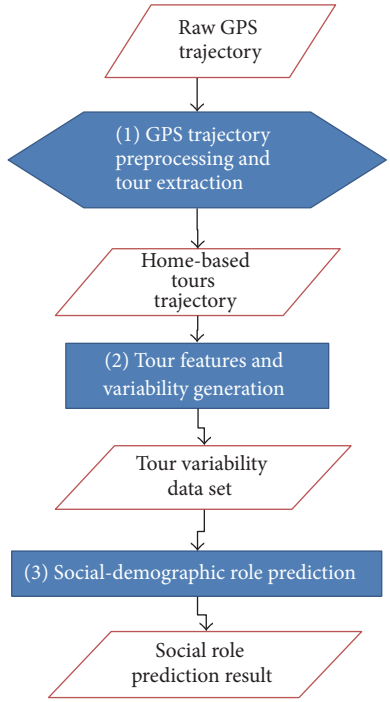

Figure 1: Methodology framework.

data collection on travel behavior variability and the social-demographic role prediction. Finally, the Conclusion closes the paper, and the principal findings are illustrated.

2. Methodology

The framework of the proposed social-demographic role prediction method is shown in Figure 1. The framework includes three modules: (1) GPS trajectory preprocessing and tour extraction, (2) tour features and variability generation, and (3) social-demographic role prediction. First, it preprocesses the raw GPS data and determines travelers' home locations. Based on that, the home-based tours from the original trajectories are detected. Then, the method extracts the tour variability by the tour features and variability generation procedure. Next, the individual tour variability data set is fed into the social role prediction module to estimate social-demographic roles. The methodology details are illustrated in the following sections.

2.1. GPS Trajectory Preprocessing and Tour Extraction. The GPS trajectory preprocessing aims to remove outliers from vehicle-instrumented GPS data. Initially, a data cleaning and smoothing process derived from Schüssler's raw GPS data-processing procedure [38] is carried out to address GPS system errors, such as warm/cold start problems, and random errors, such as urban canyon errors. Several criteria are used for removing system errors, like the number of satellites, the ground elevation, and the distance between consecutive GPS points. A procedure to detect repeated measurements will record nearly the same coordinates and zero or almost zero travel speed measurements for two or more consecutive GPS points. Only one point represents the repeated measurements: for example, a vehicle stopped at a location in front of a red light will only be represented by one GPS point rather than duplicated measurement points. After the data cleaning and filtering processes have been applied, most of the outliers will be removed, and GPS data trajectories are ready for use.

For continuous GPS trajectory data, it is not hard to detect the home location and then to generate home-based tours. The top three most visited places clusters' centroid location that the user has visited (departure from or arrive to) are at least 1 mile away from each other, as they are more likely to be home or other locations, such as the workplace. According to the characteristics of the trips related to these sites, such as the departure location of the first trip of the day, the arrival location of the last trip of the day, and the duration of the stay (e.g., more than 8 hours) at this site, home locations can be identified.

After determining the home location, the individual home-based tours, such as Home-to-Home (HH) and Home-to-Non-Home (HN), can be produced. An HH tour is defined as the traveler departing from and returning back home with a reasonable trip travel time during the day (such as 3 hours). An HN tour is the travel during the day departing from home and arriving at any other location, such as the workplace. In a day, the HH and HN tour number, especially the HH tour number, may be greater than 1.

2.2. Tour Features and Variability Generation. In this study, a home-based tour, either HH or HN, may comprise one or more consecutive trips, which is described by departure time, destination location, driving time, and travel time. Similar to the trip, a tour has the tour features encompassing departure time, destination location, driving time ratio, and travel time. The *tour departure time* is the first trip departure time of the tour, which is a temporal travel behavior feature. The spatial feature, *tour destination location*, includes the in-tour trips destination locations and the tour destination location, which is represented by a position coordinates set $\{d_1 = (\text{lat}_1, \text{lon}_1), d_2 = (\text{lat}_2, \text{lon}_2), \ldots\}$. The *tour travel time* is defined as the total elapsed time (in minutes) from the tour origin (i.e., home) to the destination (i.e., home or others), while the *tour driving time ratio* is calculated by all trips' driving time over the tour travel time. Both *tour travel time* and *tour driving time ratio* are mobile features or "degree of trip chain." The travel behavior variability is derived from the variance pattern of tour feature during the data collection period.

For a traveler i at jth tour, $x_{i,j}^{t,k}$ represents the tour feature of home-based tour ($t = \text{HH}, \text{HN}$, $k = 1, 2, 3, 4$ [1-departure time; 2-destination location; 3-travel time; 4-driving time ratio]), and $v_{i,n}^{t,k}$ denotes the tour variability of different tour features, derived from n tours. The descriptions of tour features and variability variables are listed in Table 1. The details of the tour features and variability are elaborated in the following parts.

Table 1: Tour feature and variability variables description.

Tour type	Features	Variability
HH	Departure time	Departure time SEM*
	Destination locations	Destination locations entropy
	Travel time	Travel time SEM
	Driving time ratio	Driving time ratio SEM
HN	Departure time	Departure time SEM
	Destination locations	Destination locations entropy
	Travel time	Travel time SEM
	Driving time ratio	Driving time ratio SEM

*SEM: expectation of standard error of the sample mean.

2.2.1. Temporal Feature and Variability. The tour temporal feature is the tour departure time, which is converted into a 15-minute time slot index from the beginning (00:00 a.m.) of the day, to describe the departure time within a day numerically. In that case, any time of day can be expressed as the 15-minute time slot index integer ranging from 0 to 95. The tour temporal variability $v_{i,n}^{t,1}$ is defined as the expectation of standard error of the sample mean (SEM) of departure time slot index $x_{i,j}^{t,1}$ crossing n tours in type t. The variability of departure time feature $v_{i,n}^{t,1}$ (i.e., departure time SEM) is illustrated below, where $\overline{x_{i,j}^{t,1}}$ is the sample mean of $x_{i,j}^{t,1}$

$$v_{i,n}^{t,1} = \sqrt{\frac{\sum_{j=1}^{n}\left(x_{i,j}^{t,1} - \overline{x_{i,j}^{t,1}}\right)^2}{n(n-1)}}. \quad (1)$$

2.2.2. Spatial Feature and Variability. The spatial feature is represented by the destination locations, which are the destinations of all trips in the tour. For example, although the HH tours have a fixed origin and destination (i.e., home), an HH tour may include multiple trips with different purposes, such as grocery shopping trips, children-pickup trips, or social trips. They may have different destination locations. For HN tours, except for the tour destination variation, the in-tour trip destination locations may vary a lot like the HH tours. To numerically describe the variability of the destination locations, Shannon's entropy [34, 39] is used in this study.

First, for individual i and tour type t, all destination locations from $x_{i,j}^{t,2}$ for n tours are collected. The total destination locations of individual i for n tours are represented as a random variable $D_i^t = \{d_{i,j}^t,\ j = 1, 2, \ldots, m\}$, $m = \sum_j \|x_{i,j}^{t,2}\|$, where the m locations are denoted as $d_{i,j}^t$, and $m \geq n$ because a tour may have more than one trip. A clustering procedure merges the close destination locations into clusters according to the distance between any two locations less than 1 km. After the location merging and clustering procedure, the clusters' centroid locations for a traveler are collected as $D'^t_i = \{d'^t_{i,j},\ j = 1, 2, \ldots, m'\}$, $m' \leq m$. Location variability of individual i for n tours can be measured as the entropy below,

$$v_{i,n}^{t,2}\left(D'^t_i\right) = -\sum_j P\left(D'^t_i = d'^t_{i,j}\right)\log_2 P\left(D'^t_i = d'^t_{i,j}\right), \quad (2)$$

where $P(D'^t_i = d'^t_{i,j})$ is the historical probability of individual i's visiting the clustering location $d'^t_{i,j}$ during n tours for tour type t. The property of Shannon entropy indicates that if a traveler repeatedly visits a single location, the location variability of the individual equals zero, while a larger value of $v_{i,n}^{t,2}$ results from regular visits to a larger number of locations.

2.2.3. Mobile Features and Variability. The mobile features reflect the vehicle movement behavior and travel property. They are delineated by travel time and driving time ratio. The variability of tour travel time $v_{i,n}^{t,3}$ is defined as the SEM of the tour travel time $x_{i,j}^{t,3}$ for n tours. Similar to the travel time, the driving time ratio variability $v_{i,n}^{t,4}$ is calculated by the SEM of the tour driving time ratio $x_{i,j}^{t,4}$ for n tours. The details are illustrated by (3), where the notations are similar as the previous section

$$v_{i,n}^{t,3} = \sqrt{\frac{\sum_{j=1}^{n}\left(x_{i,j}^{t,3} - \overline{x_{i,j}^{t,3}}\right)^2}{n(n-1)}},$$
$$v_{i,n}^{t,4} = \sqrt{\frac{\sum_{j=1}^{n}\left(x_{i,j}^{t,4} - \overline{x_{i,j}^{t,4}}\right)^2}{n(n-1)}}. \quad (3)$$

2.3. Social-Demographic Role Prediction. After collecting individuals' variability variables, with the individuals' social-demographic role labels as the ground truth data, a supervised machine learning model describing the correlation between travel behavior variability and social-demographic role can be developed. The eight variability variables are the independent features for defining an individual's travel behavior variability pattern, and the ground truth social-demographic role is used as the dependent variable. The support vector machine (SVM) [36, 40] is a favorite and the most used supervised machine learning approach for multiple and binary classifications and prediction applications. SVM is known as a large margin classifier, and it determines the best decision hyperplanes that provide the biggest possible margin among classes. The primal problem is formatted as

$$\min_{w,b,\varepsilon}\ \left(\frac{1}{2}W^T W + C\sum_{i=1}^{n}\varepsilon_i\right)$$
$$\text{subjected to:}\quad y_i\left(W^T\phi(x_i) + b\right) \geq 1 - \varepsilon_i,\quad \forall i \quad (4)$$
$$\varepsilon_i \geq 0,\ \forall i,$$

where W is the weight vector of features to define the decision boundary; $C\sum_{i=1}^{n}\varepsilon_i$ is a regularization (or penalty) term to

relax the objective function, where ε_i is the distance of the point from the margin if it is misclassified, and C is a constant coefficient to weight the penalty; b is the intercept and $\phi(x_i)$ is the data transformation function; n represents the data sample size; and y_i is the class label for data sample i (i.e., −1 or 1 for binary classes). The dual problem is developed to help in solving the constrained optimization primal problem,

$$\begin{aligned} \max \quad & \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi^T(x_i) \phi(x_j) \right) \\ \text{subjected to:} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \forall i \\ & 0 \le \alpha_i \le C, \ \forall i, \end{aligned} \tag{5}$$

where α_i is the Lagrange multiplier, which is the decision variable. The dual objective function can be represented by the kernel $K(x_i, x_j) = \phi^T(x_i)\phi(x_j)$. The radial basis function kernel was suggested to be the most appropriate kernel [41, 42] and was used in this model. The dual problem solutions, which are Lagrange multiplier α_i, are used for predicting the data class by computing the decision function $f(x) = W^T\phi(x) + b = \sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b$, where m is the vector dimension number (eight travel behavior variability variables in this study). The binary classification is determined by the positive or negative values of the decision function $f(x)$.

3. Case Study and Discussion

The Puget Sound Regional Council traffic choices study was an 18-month (during 2004 to 2006) research on travel behavior in response to road use. With 450 vehicles from over 275 households, the GPS raw trajectory data indicated that more than 4.5 million vehicle miles were traveled. Travelers' social-demographic features are collected as well. The National Renewable Energy Laboratory's Transportation Secure Data Center [37] summarized the data with high-resolution GPS trajectory data and traditional household survey data. In this experiment, the home-based tours features are extracted from the raw GPS data and, based on that, the variability of tour features is generated. In conjunction with the collected individual social role data, taking tour variability features as the independent variables, an SVM-based prediction model is developed and validated. The number of tours the threshold sensitivity analysis presented based on the experiment data indicates how the thresholds impact tour variability and prediction.

3.1. Experiment

3.1.1. Case Study and Variability Observations. After the raw data were preprocessed and incomplete records were removed, a total of 218 individuals have complete variability variables for at least five HH or HN tours with social-demographic information. For those 218 individuals, the individual's HH tours (green) and HN tours (red) number distributions are illustrated Figure 2. The mean value of HH tours is about 195, while the average value of HN tours is about 155. One observation is that the HH histogram shifted toward the right-hand side, which implies that there are more HH tours than there are HN tours for general travelers. That is because there are more HH tours than HN tours during a day.

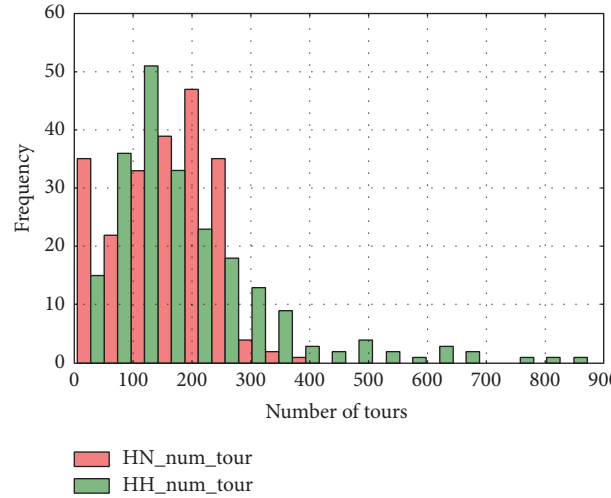

Figure 2: Histogram of number of HH and HN tours.

The individuals' social-demographic roles (employment status) include six types: (1) full-time employee; (2) part-time employee; (3) student; (4) homemaker; (5) retired; and (6) other. The number of type 1-full-time employees dominates the other types. Considering the unbalanced data amount of social role types, the original data set is converted as a binary class data set as type 1 and type 0. Type 1 class is the original type 1 class, while type 0 class stands for the total of type 2 through type 6. Type 1 class has 165 travelers; type 0 class has 53 travelers. The tours' variability variables of the binary class data set are discussed. Table 2 illustrates the statistical details for type 1 and type 0.

The statistically significant variables are *HH tours departure time SEM and HN tours departure time SEM and HN tours driving time ratio SEM*.

(i) For HN tours, type 1 travelers have significantly lower mean values of *departure time SEM* (2.51 versus 4.32) than that of type 0 travelers. The reason behind is that type 1 travelers have more departure time restriction on home to other places tours, for example, morning home-to-work commute.

(ii) For HN tours, the HN tour's mean *driving time ratio SEM* for type 1 travelers is smaller than that for type 0 (0.06 versus 0.08), which indicates that driving time ratio change of type 1 is not significant as that of type 0. It can be explained as the type 1 travelers are more dedicated to their trips and do not frequently stop during their tours.

(iii) For HH tours, the departure time situation is reversed. The mean *departure time SEM* of type 1 is 6.81, which is higher than that of type 0 (5.84). It indicates that the type 1 travelers have slightly more departure time variability for HH tours.

Table 2: Variability variables statistical details for binary class.

Variability variables	Type 1		Type 0			
	Mean	Std.	Mean	Std.	t-value	p value
HH						
Departure time SEM**	6.81	2.53	5.84	2.24	2.64	0.0096
Destination locations entropy	1.98	0.64	2.11	0.68	−1.26	0.213
Travel time SEM	62.76	48.34	54.59	42.06	1.09	0.28
Driving time ratio SEM	0.1	0.05	0.1	0.05	−0.77	0.445
HN						
Departure time SEM***	2.51	2.81	4.32	3.19	−3.66	0.0004
Destination locations entropy	0.89	0.74	1.06	0.88	−1.32	0.19
Travel time SEM	23.64	47.3	27.61	50.61	−0.5	0.621
Driving time ratio SEM*	0.06	0.06	0.08	0.07	−2.24	0.028

*Significant at level of 0.05. **Significant at level of 0.01. ***Significant at level of 0.001.

Table 3: Multiclass and binary class employment status SVM prediction results.

(a)

Employment status-multiclass		Estimation							Recall accuracy
	Type	1	2	3	4	5	6	total	
Actual	1	165	0	0	0	0	0	165	100.00%
	2	3	20	0	0	0	0	23	86.96%
	3	3	0	3	0	0	0	6	50.00%
	4	2	0	0	14	0	0	16	87.50%
	5	1	0	0	0	2	0	3	66.67%
	6	2	0	0	0	0	3	5	60.00%
	Total	176	20	3	14	2	3	218	—
	Precision accuracy	93.75%	100%	100%	100%	100%	100%	—	94.95%

(b)

Employment status-binary class		Estimation			Recall accuracy
	Type	1	0	Total	
Actual	1	165	0	165	100.00%
	0	11	42	53	79.25%
	Total	176	42	218	—
	Precision accuracy	93.75%	100%	—	94.95%

3.1.2. Social-Demographic Role Prediction Result. In the prediction model, the SVM classification is implemented by the python library (sklearn) taking default configurations, and radial basis function kernel is used. The multiclass and binary class prediction accuracy results are illustrated in Table 3. It lists two accuracy metrics. The *recall accuracy* is defined as the correctly estimated individuals' number over the total number of actual individuals of the type class. The *precision accuracy* is the ratio of correctly estimated individuals' number over the total number of estimated individuals of the type class.

From Table 3, the prediction results are promising, and the overall general accuracy of prediction reaches 94.95%. For multiclass prediction, type 1 class has the highest recall accuracy (100%), type 3 student class has the worst recall accuracy (50%), and three of them are falsely labeled as full-time employees. From a precision accuracy perspective, all classes have high precision accuracy values. For binary class prediction, the recall of type 1 class is still 100%, and 11 travelers from type 0 class are predicted as type 1, which generates a recall of 79.25%. The prediction accuracies for type 1 and type 0 are 93.75% and 100%, respectively.

Table 4: Income level multiclass SVM prediction results.

Income level-multiclass	Estimation						
	<$25K	$25K–50K	$50K–75K	$75K–150K	>$150K	Total	Recall accuracy
Actual							
<$25K	16	0	1	3	0	20	80.00%
$25K–50K	0	26	0	10	0	36	72.22%
$50K–75K	0	0	33	8	0	41	80.49%
$75K–150K	0	0	1	101	0	102	99.02%
>$150K	0	0	1	5	13	19	68.42%
Total	16	26	36	127	13	218	—
Precision accuracy	100.00%	100.00%	91.67%	79.53%	100.00%	—	**86.70%**

Table 5: Age level multiclass SVM prediction results.

Age level-multiclass	Estimation							Recall accuracy
	<21	22–34	35–44	45–54	55–65	>65	Total	
Actual								
<21	0	1	0	2	0	0	3	0.00%
22–34	0	37	0	7	0	0	44	84.09%
35–44	0	1	51	11	0	0	63	80.95%
45–54	0	3	0	71	0	0	74	95.95%
55–65	0	1	1	4	19	0	25	76.00%
>65	0	0	1	5	0	3	9	33.33%
Total	0	43	53	100	19	3	218	—
Precision accuracy	—	86.05%	96.23%	71.00%	100.00%	100.00%	—	**83.03%**

Table 6: Gender multiclass SVM prediction results.

Gender status-binary class	Estimation			Recall accuracy
	Female	Male	Total	
Actual				
Female	130	5	135	96.30%
Male	15	68	83	81.93%
Total	145	73	218	—
Precision accuracy	89.66%	93.15%	—	**90.83%**

One observation of the results is the poor prediction performance of type 2 to type 6 classes in the multiclass case and type 0 in binary class cases. The poor prediction results may be led by the unbalanced data set and the limited sample size.

3.1.3. Income, Age, and Gender Prediction Results. In addition to the employment status, an individual's other social-demographic variables, including income, age, and gender, are discussed in this study. Similar to the experiment results of employment status shown previously, the prediction results of those three variables (income, age, and gender) are shown in Tables 4, 5, and 6. The individual's income is defined at five levels: (1) less than $25,000, (2) $25,000–$50,000, (3) $50,000–$75,000, (4) $75,000–$150,000, and (5) greater than $150,000. The individual's age is categorized in different classes: (1) less than 21, (2) 22–34, (3) 35–44, (4) 45–54, (5) 55–65, and (6) greater than 65. The individual's gender includes female and male.

The overall prediction accuracy values of the three variables (income level = 86.7%, age level = 83.03%, and gender = 90.83%) are still acceptable, although they are relatively lower than the prediction accuracy of employment status (94.95%). It indicates that individual's employment status is easier to predict than other variables. The reason behind is that the employment status is more directly and closely correlated to the travel behavior variability than other social-demographic variables.

3.2. Sensitivity Analysis. The test data were collected over nearly 18 months, and for a data set collected over a long time, it is feasible to carry out a sensitivity analysis for the sampling threshold, that is, the number of tours. The sensitivity analysis investigates how the threshold impacts the tour variability and even social-demographic role prediction, aiming to answer the questions about the data collection sufficiency for travel behavior variability convergence and estimating the individuals' social-demographic roles. As a comparison to the SVM model used in the study, another machine learning classification model, logistic regression (LR), is implemented in the analysis.

The number of tours threshold is defined as the required minimum number of tours for both HH and HN for a successful data collection. The number of tours threshold ranges as [1, 2, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 180]. For example, a value of 5 indicates that

any travelers with less than five tours are disqualified and discarded, while the travelers with 5 or more are qualified and collected. For the qualified individuals, five tours are randomly selected from this traveler's tour pool. If the threshold value is high, the number of individuals satisfying the number of tours requirement becomes small and vice versa.

3.2.1. Tour Variability. The tour variability variables plotted against the number of tours thresholds for type 1 and type 0 travelers are illustrated in Figure 3. For all diagrams, the x-axis is the number of tours thresholds, and the y-axis is the variability variables; each single curve represents an individual's variability values along with the number of tours thresholds. In the diagrams, all variability variables of type 1 and type 0 follow the *same or similar patterns*.

The HH and HN tours' *departure time SEM, travel time SEM, and driving time ratio SEM* variability curves are oscillating at the beginning (lower number of tours thresholds of 40) and then converge to the small values at the end. This indicates that a larger sample size will reduce the variability of the tour features.

The *destination location entropy* of HH tours (Figures 3(c) and 3(d)) dramatically increases at the beginning (at about 40 tours) and then converges to large individual values for all people. This means that more samples will bring more uncertainty to the destination locations at a small threshold range and then stays constant for the high thresholds. However, the destination location entropy of HN tours (Figures 3(k) and 3(l)) does not follow the significant increase and convergence pattern.

Generally, for a large sample size, the variances of travel behavior features will not change too much, and the variability values are low. According to the diagrams, one thumb of rule is that when the number of tours reaches about 40, the variances of travel behavior features keep constant at low values (except destination location entropy) and the travel behavior variability is more reliable and predictable.

The statistical analyses of two types of travelers for all eight travel behavior variability variables are conducted to understand the travel behavior features variances "before and after 40 tour threshold." The statistical results are listed in Table 7. The feature variability values for each type of traveler are separated into two groups: "equal to and less than 40" (≤40) and "greater than 40" (>40), according to the number of tours threshold attributes. The sample sizes of the two groups are comparable for each type of travelers. For type 0, the "≤40" group has 379 measurements, while the ">40" group has 197 measurements. For type 1, the "≤40" group has 1,272 measurements, while the ">40" group has 847 measurements.

The *standard deviation* and *mean values* of the ">40" group for each type of traveler at nearly all features are significantly smaller than those of the other group ("≤40"), except for a few cases (e.g., *HN location entropy* and *HN travel time SEM* at type 1). Besides, for almost all cases, hypothesis tests are significant, except the *HN travel time SEM* at type 1 and *HH travel time SEM* at type 0. This indicates the two groups are statistically different for two types of traveler.

The statistical analysis results are consistent with the observations from Figure 3. They validate the conclusion that when the number of tours is more than 40, the variances of travel behavior features keep constant at low values (except for destination location entropies, which are at high values) compared to the cases which are within the "equal to and less than 40" group.

3.2.2. Social-Demographic Role Prediction. The sensitivity study includes logistic regression (LR) as a comparable prediction approach to the SVM model used in this study. This comparison study focuses on the data set overall recall accuracy. Since the number of qualified individuals decreases as the number of tours threshold goes up, the various sample set sizes at different thresholds may impact the prediction results. Figure 4 describes decreasing trend of the number of qualified individuals as the number of tours thresholds increases.

From Figure 4, we can see that, after 90, the decreasing trend of the number of qualified individuals is more significant. The number of qualified individuals at thresholds after 90 is almost less than 100. A fixed sample set, which includes the 126 qualified individuals at threshold 90 (who have at least 90 tours), is used for the test in the study. Those 126 qualified individuals exist as a subset in the qualified individual sets at thresholds from 5 to 90.

The sensitivity research result for the fixed sample set for the number of threshold ranging from 5 to 90 is illustrated in Figure 5. The SVM and LR prediction accuracy values roughly keep constant throughout all different thresholds, while the average prediction accuracy of SVM (about 95%) is always better than that of the LR model (about 84%).

The prediction results illustrate that a larger number of tours required for data collection does not significantly improve the prediction accuracy. Since the traveler type detection result heavily depends on the travel behavior variability differences between both types of travelers, the same or similar travel behavior variability patterns of both types of individuals (which are observed from the diagrams in Figure 3) may explain the result. Although most travel behavior features' variability converges as the number of tours threshold increases, the relative variability difference of type 1 and type 0 travelers may not change much for different thresholds. Also, for the fixed sample set (126 qualified individuals), the two types of individuals' variability mean difference ratio ((type_1 − type_0)/type_1) of each travel behavior variability variable, crossing different thresholds, are used to describe the relative variability difference indirectly and to help understand the prediction result, statistically. Figure 6 illustrates the variability mean difference ratio changing trend of all eight variability features as the number of tours threshold increases. From it, the variability mean difference ratios of most features (except *(3) HH travel time SEM* and *(7) HN travel time SEM*) stay low and keep constant crossing all thresholds. It tells that the relative variability differences of almost all features are not significantly changed by different thresholds. It partially explains that increase in the number of tours threshold does not cause significant

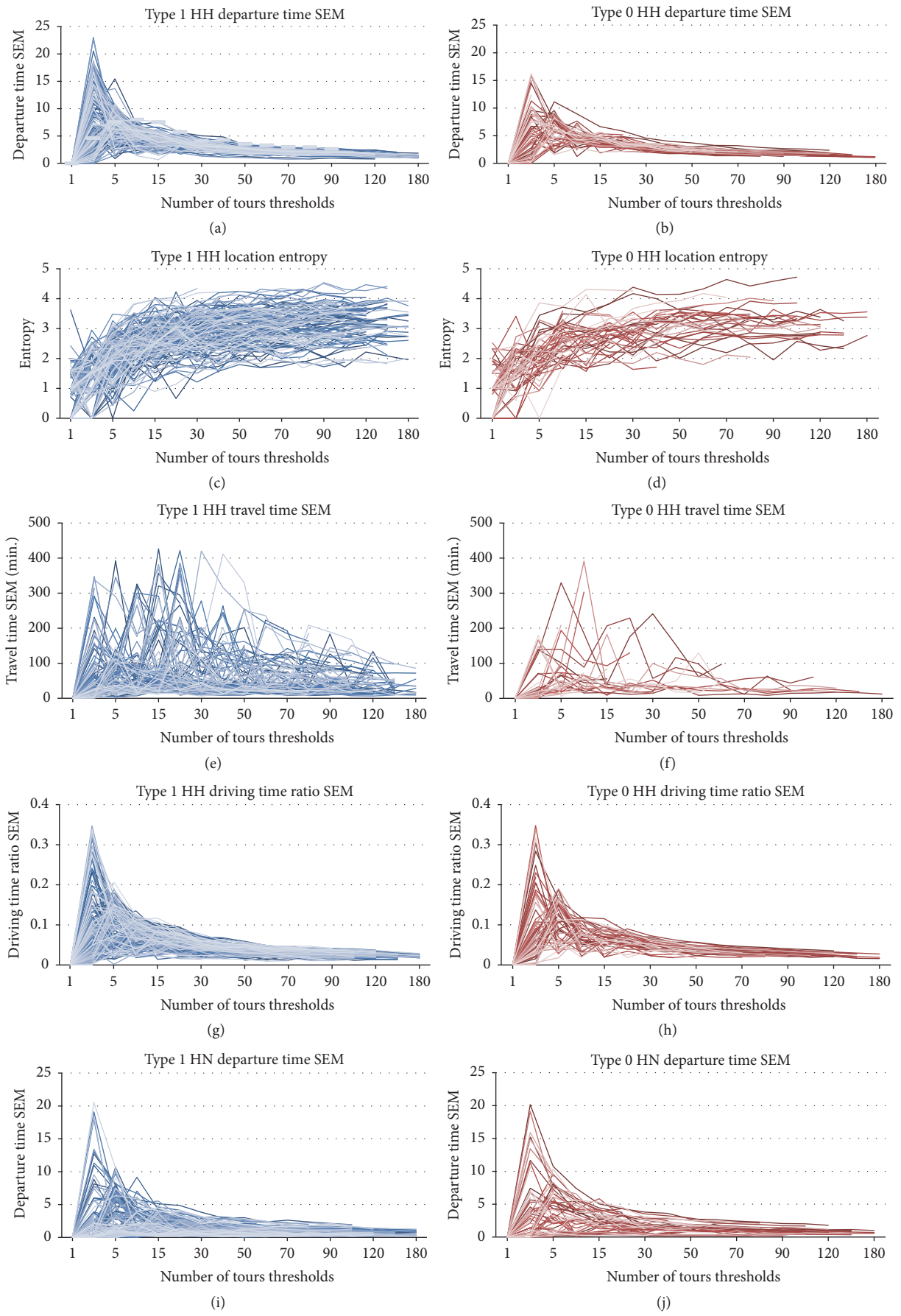

Figure 3: Continued.

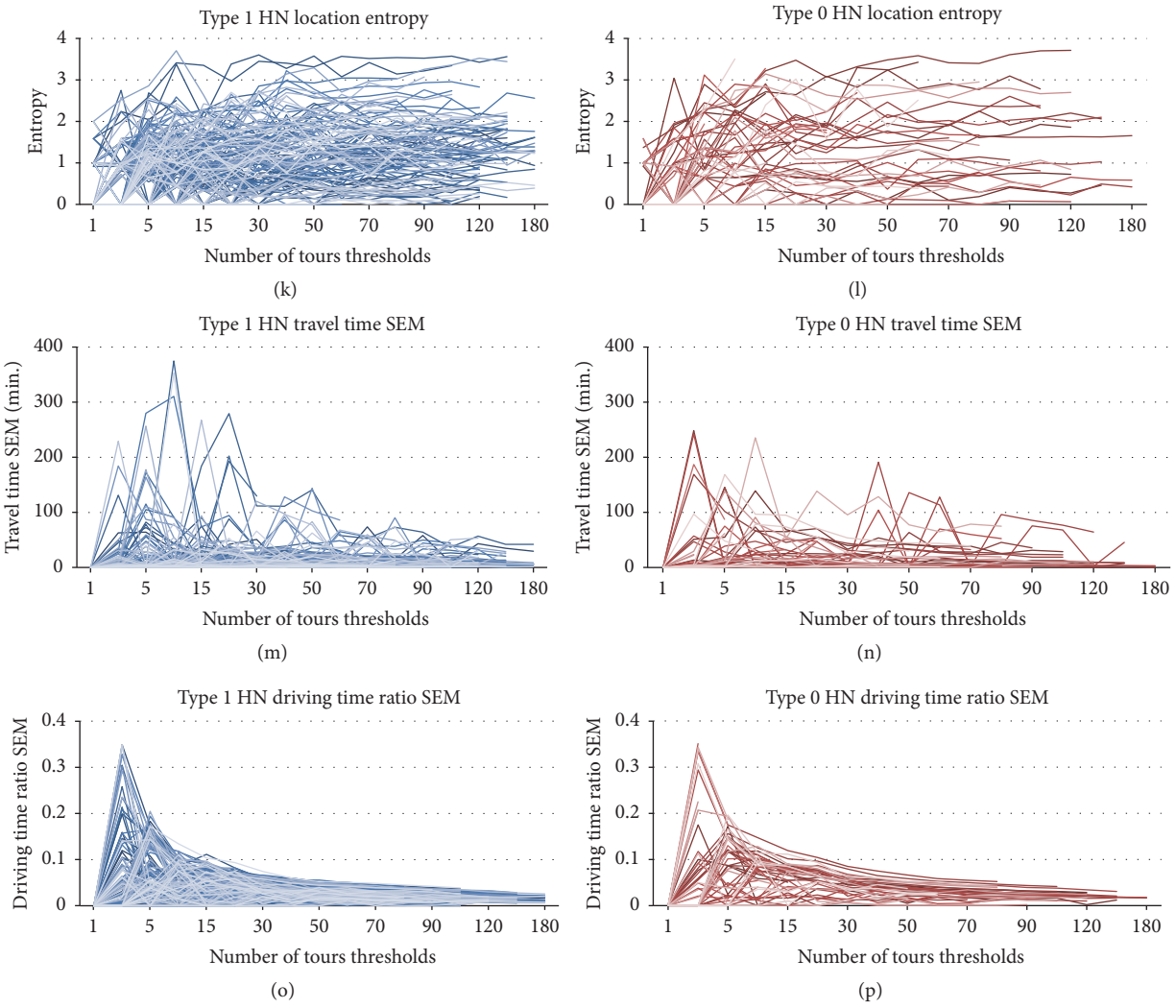

Figure 3: Variability versus number of tours thresholds.

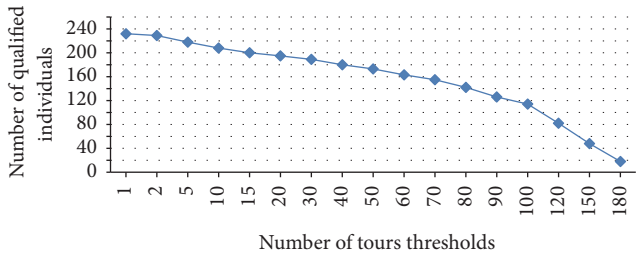

Figure 4: Number of qualified individuals versus number of tours thresholds.

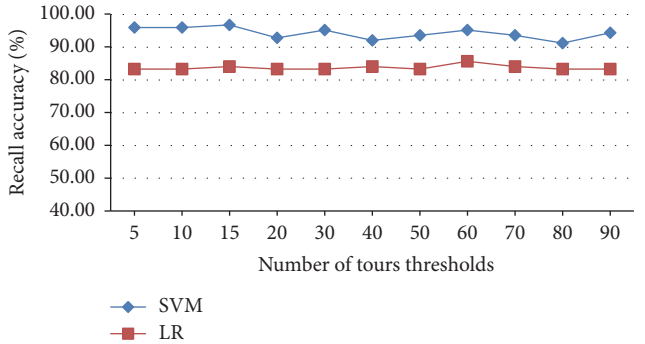

Figure 5: Sensitivity study of prediction accuracy and number of tours threshold for a fixed sample set.

social-demographic prediction changes, at least for this fixed sample set.

4. Conclusions

This paper proposes a social-demographic role prediction method based on the travel behavior variability pattern. It is based on the principles that, for different social groups, they have specific travel behavior patterns. The paper provides a way to formalize traveler's travel behavior variability pattern by analyzing long-term raw GPS data and to predict

Table 7: Statistical analysis of variances of travel behavior features for threshold-40.

Features	Statistical measures	Type 1		Type 0	
		≤40 (1272)	>40 (847)	≤40 (379)	>40 (174)
HH departure time SEM	Mean	4.17	1.99	3.66	1.94
	Std.	3.28	0.46	2.90	0.47
	t-value	23.26	—	11.22	—
	p value	0.00	—	0.00	—
HH location entropy	Mean	**2.21**	**3.10**	**2.20**	**3.15**
	Std.	0.92	0.57	0.95	0.51
	t-value	−27.52	—	−15.08	—
	p value	0.00	—	0.00	—
HH travel time SEM	Mean	129.84	103.76	393.90	164.91
	Std.	388.96	164.87	409.97	264.37
	t-value	2.12	—	1.82	—
	p value	0.03	—	**0.07**	—
HH driving time ratio SEM	Mean	0.06	0.03	0.06	0.03
	Std.	0.05	0.01	0.06	0.01
	t-value	21.29	—	10.22	—
	p value	0.00	—	0.00	—
HN departure time SEM	Mean	1.46	0.83	2.42	1.11
	Std.	2.23	0.59	2.95	0.60
	t-value	9.53	—	8.22	—
	p value	0.00	—	0.00	—
HN location entropy	Mean	**0.89**	**1.32**	**0.98**	**1.38**
	Std.	0.77	0.75	0.93	0.98
	t-value	−12.89	—	−4.44	—
	p value	0.00	—	0.00	—
HN travel time SEM	Mean	**46.38**	**52.15**	36.16	23.12
	Std.	276.03	192.73	103.15	37.95
	t-value	−0.57	—	2.16	—
	p value	**0.57**	—	0.03	—
HN driving time ratio SEM	Mean	0.04	0.02	0.05	0.03
	Std.	0.05	0.01	0.06	0.01
	t-value	11.24	—	7.34	—
	p value	0.00	—	0.00	—

individuals social-demographic roles through support vector machine model by travel behavior variability.

The study applies to Puget Sound Regional Council data set, which includes a long-term (18-month) GPS trajectory data set and a particular individual social-demographic data set. The variability derived from the data set indicates that, (1) for HN tours, the full-time employees have tighter departure time restrictions on home to other places tours, for example, the morning home-to-work commute; (2) they are more dedicated to their trips and do not stop frequently; (3) for HH tours, the full-time employee individuals have more departure time flexibility. According to the travel behavior variability properties, the prediction accuracy rates for social-demographic features, including employment status, income, age, and gender, are discovered. Among the social-demographic features, an individual's employment status is mostly related to the travel behavior variability and can be predicted accurately. The sensitivity analyses about sampling size (number of tours threshold) impacts on the tour variability and the prediction accuracy are also studied. The tour variability is going to converge as the number of tours threshold increases. However, for the fixed sample set, the social-demographic role predictions do not change much as the number of tours threshold increases.

This study preliminarily explores the possibility of using travel behavior variability to predict an individual's social-demographic information. This prediction method helps to obtain the social-demographic data for the people with long-term collected activity data without any traditional travel surveys. The sensitivity analysis can guide future studies to gather data and design the experiments. However, there are several limitations of this study. The first issue is that there are only a few individuals in the test data set. A larger traveler sample size may improve the model's performance: the model only considers home-based tours and limited travel behavior variability attributes. More measures of travel

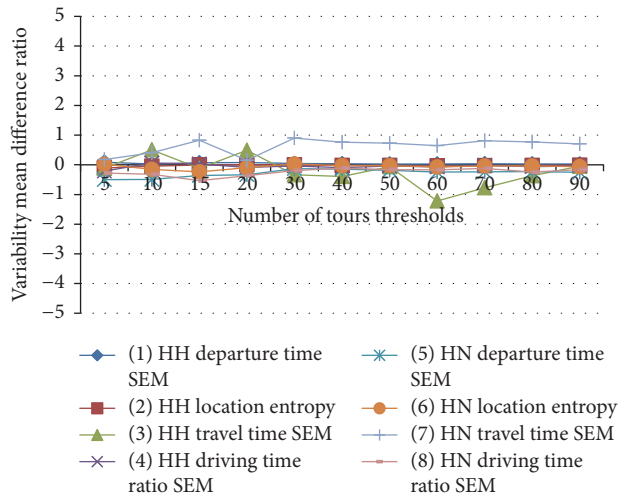

FIGURE 6: Variability mean difference ratio versus number of tours threshold.

behavior features and their variability, such as travel mode, trip purpose, and other types of tours (e.g., work-based tours), should be considered in future work.

Disclosure

The publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work or allow others to do so, for U.S. Government purposes.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Acknowledgments

This work was supported by the U.S. Department of Energy under Contract no. DE-AC36-08GO28308 with Alliance for Sustainable Energy, LLC, the Manager and Operator of the National Renewable Energy Laboratory. Funding was provided by the Federal Highway Administration.

References

[1] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, 2016.

[2] C. Chen, L. Bian, and J. Ma, "From traces to trajectories: how well can we guess activity locations from mobile phone traces?" *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 326–337, 2014.

[3] Y. Kim, F. C. Pereira, F. Zhao, A. Ghorpade, P. C. Zegras, and M. Ben-Akiva, "Activity recognition for a smartphone and web based travel survey," https://arxiv.org/abs/1502.03634.

[4] J. J. Zhou and R. Golledge, *An analysis of variability of travel behavior within one-week period based on GPS*, University of California Transportation Center, 2003.

[5] L. Zhu, J. R. Holden, and J. D. Gonder, "A trajectory segmentation map-matching approach for large-scale, high-resolution GPS data," in *Transportation Research Board 96th Annual Meeting*, Washington, DC, USA.

[6] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 363–381, 2014.

[7] X. Hu, Y. Chiu, Y. Ma, and L. Zhu, "Studying Driving Risk Factors using Multi-Source Mobile Computing Data," *International Journal of Transportation Science and Technology*, vol. 4, no. 3, pp. 295–312, 2015.

[8] T. Kusakabe and Y. Asakura, "Combination of smart card data with person trip survey data," in *Public Transport Planning with Smart Card Data*, pp. 73–92, CRC Press, 2016.

[9] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: a data fusion approach," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 179–191, 2014.

[10] X. Hu, Y. Chiu, and L. Zhu, "Behavior insights for an incentive-based active demand management platform," *International Journal of Transportation Science and Technology*, vol. 4, no. 2, pp. 119–134, 2015.

[11] A. Sharafsaleh et al., "High Occupancy Vehicle Lane Management System: Amendment A," 2011.

[12] M. Friman, L. Larhult, and T. Gärling, "An analysis of soft transport policy measures implemented in Sweden to reduce private car use," *Transportation*, vol. 40, no. 1, pp. 109–129, 2013.

[13] A. Downs, *Still Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*, Brookings Institution Press, 2005.

[14] M. W. Burris and R. M. Pendyala, "Discrete choice models of traveler participation in differential time of day pricing programs," *Transport Policy*, vol. 9, no. 3, pp. 241–251, 2002.

[15] M. W. Burris, "Application of variable tools on congested toll road," *Journal of Transportation Engineering*, vol. 129, no. 4, pp. 354–361, 2003.

[16] Texas Transportation Institute, "Managed lanes," http://managed-lanes.tamu.edu/resources/reports.

[17] G. Pierce and D. Shoup, "Getting the prices right," *Journal of the American Planning Association*, vol. 79, no. 1, pp. 67–81, 2013.

[18] G. Möser and S. Bamberg, "The effectiveness of soft transport policy measures: A critical assessment and meta-analysis of empirical evidence," *Journal of Environmental Psychology*, vol. 28, no. 1, pp. 10–26, 2008.

[19] Y. Asakura, T. Iryo, Y. Nakajima, and T. Kusakabe, "Estimation of behavioural change of railway passengers using smart card data," *Public Transport*, vol. 4, no. 1, pp. 1–16, 2012.

[20] T. Kusakabe, T. Iryo, and Y. Asakura, "Estimation method for railway passengers' train choice behavior with smart card transaction data," *Transportation*, vol. 37, no. 5, pp. 731–749, 2010.

[21] E. Gargiulo, R. Giannantonio, E. Guercio, C. Borean, and G. Zenezini, "Dynamic ride sharing service: are users ready to adopt it?" *Procedia Manufacturing*, vol. 3, pp. 777–784, 2015.

[22] J. Goodwin, "The Future of Traffic Congestion at Your Fingertips," http://www.baycrossings.com/dispnews.php?id=1761.

[23] S. Bamberg, I. Ajzen, and P. Schmidt, "Choice of Travel Mode in the Theory of Planned Behavior: The Roles of Past

Behavior, Habit, and Reasoned Action," *Basic and Applied Social Psychology*, vol. 25, no. 3, pp. 175–187, 2003.

[24] G. Currie, "Free fare incentives to shift rail demand peaks-medium-term impacts," *in Transportation Research Board 90th Annual Meeting*, 2011.

[25] E. Ben-Elia, H. Boeije, and D. Ettema, "Behavior Change Dynamics in Response to Rewarding Rush-Hour Avoidance: A Qualitative Research Approach," in *Transportation Research Board 90th Annual Meeting*, pp. 567–582, 2011.

[26] E. Ben-Elia and D. Ettema, "Changing commuters' behavior using rewards: A study of rush-hour avoidance," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 5, pp. 354–368, 2011.

[27] P. R. Stopher, "Use of an activity-based diary to collect household travel data," *Transportation*, vol. 19, no. 2, pp. 159–176, 1992.

[28] B. Rogers, "The social costs of uber," *University of Chicago Law Review Dialogue*, vol. 82, no. 3, pp. 85–103, 2015.

[29] S. Hanson and J. O. Huff, "Assessing day-to-day variability in complex travel patterns," 1981.

[30] J. O. Huff and S. Hanson, "Repetition and Variability in Urban Travel," *Geographical Analysis*, vol. 18, no. 2, pp. 97–114, 1986.

[31] E. I. Pas and F. S. Koppelman, "An examination of the determinants of day-to-day variability in individuals' urban travel behavior," *Transportation*, vol. 13, no. 2, pp. 183–200, 1986.

[32] R. N. Buliung, M. J. Roorda, and T. K. Remmel, "Exploring spatial variety in patterns of activity-travel behaviour: Initial results from the Toronto Travel-Activity Panel Survey (TTAPS)," *Transportation*, vol. 35, no. 6, pp. 697–722, 2008.

[33] E. I. Pas and S. Sundar, "Intrapersonal variability in daily urban travel behavior: Some additional evidence," *Transportation*, vol. 22, no. 2, pp. 135–150, 1995.

[34] M. Wang, C. Chen, and J. Ma, "Time-of-day dependence of location variability: application of passively-generated mobile phone data set," in *Transportation Research Board 94th Annual Meeting*, 2015.

[35] M. Allahviranloo, "Pattern Recognition and Personal Travel Behavior," *in Transportation Research Board 95th Annual Meeting*, 2016.

[36] M. Allahviranloo and W. Recker, "Daily activity pattern recognition by using support vector machines with multiple classes," *Transportation Research Part B: Methodological*, vol. 58, pp. 16–43, 2013.

[37] Transportation Secure Data Center, "National Renewable Energy Laboratory," https://www.nrel.gov/transportation/secure-transportation-data.html.

[38] N. Schüssler, *Accounting for Similarities between Alternatives in Discrete Choice Models Based on High-Resolution Observations of Transport Behaviour*, ETH Zurich, 2016.

[39] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[40] C. M. Bishop, "Pattern recognition," 2006.

[41] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, 2003.

[42] L. Lin, Q. Wang, S. Huang, and A. W. Sadek, "On-line prediction of border crossing traffic using an enhanced Spinning Network method," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 158–173, 2014.