IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2020, accepted December 29, 2020, date of publication January 1, 2021, date of current version January 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048847

Inferring Inter-City Trip Purpose From the Perspective of the Group

JIANPEI QIAN, CHUNFU SHAO, CHUNJIAO DONG, AND SHICHEN HUANG
Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Ministry of Transport, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Chunfu Shao (cfshao@bjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 71621001.

ABSTRACT Although trip purpose inference based on passively collected data has long been investigated, less attention has been paid to inter-city trips. The reason is, except using ticket sales data, only limited trips can be extracted due to the lower frequency of inter-city trips during daily life. However, for ticket sales data, only limited features can be explored due to the lower spatial resolution of trajectories. Therefore, this paper endeavoured to exploit the potential of ticket sales data from the perspective of the group. Theoretically, by introducing concepts of text mining, the trip purpose of a group can be viewed as analogous to the topics of a document. Trip purpose was characterized by a time topic model (TTM) that incorporates start time, in contrast to latent Dirichlet allocation (LDA). This approach was implemented via a three-step method. First, groups were reconstructed from tickets. Second, three types of features, i.e., demographic, experience and co-travel network features, were extracted as a series of words to describe passengers. Third, trip purposes were automatically clustered based on the co-occurrence of words in the same group using a TTM. This paper presents comparison experiments to evaluate feature sets and the model performance based on a web-based travel survey, including the ground truth. Moreover, this paper highlights the practical use of a TTM to detect anomalies beyond anticipated trip purpose based on large-scale ticket sales data collected from Beijing, China. The full feature set was found to be preferable since both *precision* and *recall* increased when demographic and co-travel network features were considered. Meanwhile, the TTM produced robust and balanced predictions and exhibited additional power to recognize personal business compared with baseline methods.

INDEX TERMS Inter-city trip, ticket sales data, topic model, trip purpose inference.

I. INTRODUCTION

Passively collected data, obtained as important supplements of travel surveys, have received considerable attention, as they ease the burden of respondents and provide accurate and massive data [1]. However, trip purpose information is usually missing, making such data intractable for relevant operators to offer further personalized services. Accordingly, researchers have investigated several methods to infer trip purpose [2], especially based on passively collected data that originate from location-based service (LBS), e.g., call details records (CDR) [3], check-in records on social media [4] and global positioning system (GPS) data [5].

Intra-city trips, either a single trip or trip chains, can easily be reconstructed from LBS data as long as the locations where activities occur are identified [6]. On this basis, trip purpose is generally inferred by utilizing the detailed spatial and temporal information of stay points [4], [7].

Inter-city trips can be extracted from LBS data similarly, but the cost is extremely high due to their comparatively low frequency during daily life. Serving as the major passively collected data in inter-city transportation systems, ticket sales data are worth exploiting instead, but trip purpose remains to be inferred.

A. RESEARCH SCOPE

In contrast to LBS data, the lower spatial resolution of ticket sales data makes it difficult to recognize the real locations where activities occur. Therefore, this paper endeavours to resolve the problem of inter-city trip purpose inference from the perspective of the group.

The associate editor coordinating the review of this manuscript and approving it for publication was Rashid Mehmood.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

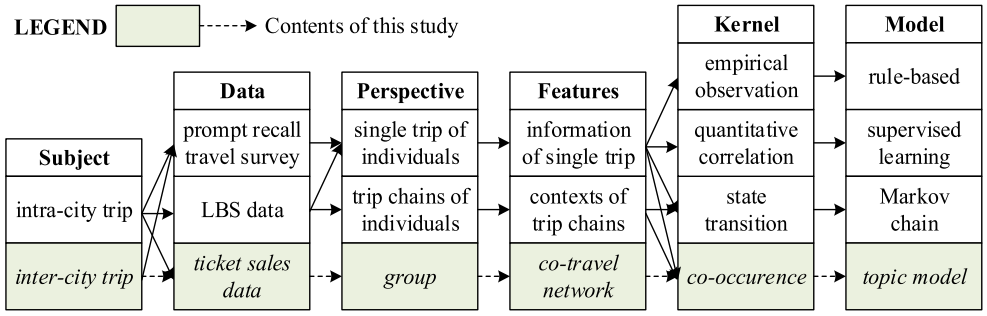

FIGURE 1. Position of this study in the research framework of trip purpose inference.

Specifically, based on a central observation that everyone has a social circle, it is natural to imagine a hidden 'circle' in inter-city trips, interpreted as a 'co-travel network' [8]. Unconsciously in many cases, people prefer to choose companions from their 'circle' according to different trip purposes, which implies that not only the features of individuals but also the composition of individuals in the same group could be potential indicators of trip purpose.

Traditionally, rule-based methods have been used to model the correlation between features and trip purpose [9]. Nevertheless, it is difficult to formulate corresponding rules for inter-city trips, because there is no circadian rhythm, as in intra-city trips [7], [10]. Subsequently, supervised learning methods have evolved as the most efficient way to establish non-linear correlations. Unfortunately, these approaches still rely on prompted recall surveys [11], [12].

Therefore, it is necessary to develop a suitable method for trip purpose inference since labelled data are often unavailable. Inspired by the *topic model,* which is widely employed in *text mining*, an analogy is made first between the concept, including the group, features of individual members and trip purpose in transportation, and the document, words and topics in the *topic model*. Then, trip purpose naturally emerges from the co-occurrence of the features in the same group.

B. MAJOR CONTRIBUTIONS

In summary, the contents of this study are depicted in **Fig. 1**, where the contributions can be generalized as twofold compared with previous studies.

The first contribution is the exploration of feature selection for inter-city trip purpose inference in light of the information contained in ticket sales data. By splitting the full set into three reduced feature sets, this study validated the improvement in model performance once demography and co-travel network features were introduced in addition to the experience features.

The second is the adoption of the *topic model* from the perspective of the group. By incorporating start time into latent Dirichlet allocation (LDA) proposed by [13], this study developed a time topic model (TTM) and compared it with baseline models.

Feature selection and model comparisons were performed based on a web-based travel survey. Eventually, the TTM was applied to large-scale ticket sales data collected from the road passenger transport system in Beijing, China, and the topics were annotated as trip purposes based on the feature distribution and start time distribution of each topic.

C. ORGANIZATION

The remainder of this paper is organized as follows. Section II reviews the previous work on feature selection and model development. Section III proposes a three-step method for feature design, group reconstruction and trip purpose generation. A Gibbs sampling algorithm for this probabilistic graphic model is also provided. Section IV conducts comparison experiments to verify candidate features and assess the model performance. Section V implements the proposed method on large-scale ticket sales data. Finally, Section VI concludes this paper.

II. LITERATURE REVIEW

A. MODEL DEVELOPMENTS

Gong *et al.* [2] classified the methods of trip purpose inference into rule-based, statistical and machine learning methods. In this paper, machine learning methods are further divided into three categories according to the extent that labelled data are required, as shown in **Table 1**.

Decision tree is the first supervised learning method to be widely adopted [14]–[16], but the *accuracy* varies greatly for different trip purposes. Thus, feature selection and even classifier selection approaches, such as bagging [17], boosting [18] and random forest [4]–[19], [20], [21], have been developed to improve the accuracy. Compared to DT, ANN achieves greater performance by balancing the *accuracy* of each trip purpose [22]. Based on the WEKA Java library, Ectors [23] compared more than ten supervised learning methods in terms of *run time* and *accuracy*. Nonetheless, these models only exploit the features discovered from a single trip. Thus, Liu *et al*. [24] proposed a post-processing algorithm to enhance the model by considering the transition probability and the prior probability between daily activity sequences, thereby achieving an increase of 7.6 percentage points.

Supervised learning methods require labelled information about trip purpose, but the labels are generalized and ambiguous. Intra-city trips are conventionally segmented into

TABLE 1. Summary of the machine learning methods in recent research on trip purpose identification.

Year	Reference	Proposed Model	Baseline models*	Accuracy**
Supervised learning				
2013	[14]	Decision tree (DT)		7.1%-96.5%
2013	[24]	Post-processing algorithm	Support vector machine (SVM), DT, logistic regression, random forest (RF)	51.4%-91.3%
2014	[19]	RF		86.2%†
2014	[15]	DT		Not available
2014	[16]	DT (based on J48 in WEKA)	Nested logit (NL)	37%-95%
2014	[20]	RF		50.0%-94.4%
2014	[17]	Ensemble DT (bagging)	k-Nearest neighbours (k-NN)	62.0%-91.0%
2015	[18]	Adaboost		12.9%-66.2%
2016	[22]	Artificial neural network (ANN)	SVM, Bayesian network (BN), multinomial logit	86.4%-99.6%
2017	[21]	Multi-stage random forest		73.2%-99.7%
2017	[4]	RF	NL	58.5%-92.2%
2017	[23]	DT (LMT, WEKA)	11 algorithms in WEKA	79.25%†
Unsupervised learning				
2016	[27]	Continuous hidden Markov model (CHMM)		
2017	[31]	Hierarchical topic modelling		
2017	[30]	LDA		
2018	[32]	POI link model (PLM)		
Semi-supervised learning				
2016	[28]	Markov random field		

* Baseline models are listed when model selection is performed in the research; ** only the accuracy of the proposed model (i.e., the recommended model) is listed; † only the average accuracy is available.

home-based and none-home-based trips according to the trip origin and destination [9]–[15] or in home, mandatory, maintenance, flexible and pick-up/drop-off according to the elasticity of start time [4], [18]. Inter-city trips are commonly categorized into business and non-business [25], [26] or tourist and business [8]. In reality, respondents may not be able to explicitly define their trip purposes according to a predefined category.

On the contrary, trip purpose can be automatically inferred using unsupervised learning methods that avoid arbitrary categorization. Han and Sohn [27] find plausible activity patterns consistent with observations by adopting CHMM. Wu and Li [28] demonstrate that the semi-supervised learning method improves the prediction of trip purpose by adding a small amount of labelled data to the training set as a penalty term.

In summary, unsupervised learning provides richer information about trip purpose via clustering. However, the prediction accuracy remains unclear. Therefore, in this paper, comparison studies are designed using a web-based travel survey to answer whether the TTM outperforms the baseline models.

B. FEATURE SELECTION

The venue information in the vicinity of the destination, especially *points of interest* (POIs), proves significant in intra-city trip purpose inference [4]–[22], [29]. Based on the idea that POIs can be regarded as words and that destinations with a variety of POIs can be regarded as documents, LDA has been successfully introduced [30], [31]. Wang *et al*. [32] extend this idea by introducing augmented O-D pairs. In other words, trip purpose is inferred from the POIs around the trip origin and destination.

The feature selection of inter-city trip purpose inference has received less attention. To the best of our knowledge, Lu and Zhang [25] are the first to address this problem. They divide potential variables into four datasets according to accessibility of the data source and design a full model, a reduced model and a minimized model to illustrate the importance of the four datasets. Notably, the data they rely on are collected from the 1995 American Travel Survey (ATS). Based on CDR, Janzen *et al*. [26] calculate the distance, duration, destination, weekend share, frequency, deviation from average distance and size of the home city to infer trip purpose. By extracting historical ticket information of passengers, Lin *et al*. [8] construct co-travel networks, in which social relations of a group are retrieved to distinguish business groups from tourist groups.

Despite the success of various works on feature selection, the low spatial resolution of ticket sales data forces us to make full use of the features in another way. Inspired by the concept of co-travel networks introduced in [8] and the observation that passengers are inclined to travel together on inter-city trips, this paper considers features from the perspective of the group. In this sense, trip purpose is inferred from the co-occurrence of the features of individual members in a group.

III. METHODOLOGY

A. RESEARCH FRAMEWORK AND GENERATION OF TRIP PURPOSE

The hypotheses for model development are as follows:

- H-1: There exists a multinomial distribution of trip purpose with a Dirichlet prior for each group;
- H-2: There exists a multinomial distribution of features that jointly portray a passenger with a Dirichlet prior for each trip purpose;

TABLE 2. List of notations.

Symbol	Definition
z	A latent variable of trip purpose
w	Observed variables of features
t	An observed variable of start time
$\mathcal{Z}$	Data set of z
$\mathcal{W}$	Data set of w
$\mathcal{T}$	Data set of t
m	A suffix denoting a certain group
n	A suffix denoting a certain w in m
k	A suffix denoting the value of z
v	A suffix denoting the value of w
l	A suffix denoting the value of t
M	A constant denoting the number of groups
N	A constant denoting the count of w, t in k or k in m
K	A constant denoting the number of values of z
V	A constant denoting the number of values of w
L	A constant denoting the number of values of t
$\boldsymbol{\theta}$	An M-dimensional multinomial distribution of z
$\boldsymbol{\varphi}$	A K-dimensional multinomial distribution of w
$\boldsymbol{\psi}$	A K-dimensional multinomial distribution of t
α	A symmetric Dirichlet prior parameter of $\boldsymbol{\theta}$
β	A symmetric Dirichlet prior parameter of $\boldsymbol{\varphi}$
γ	A symmetric Dirichlet prior parameter of $\boldsymbol{\psi}$

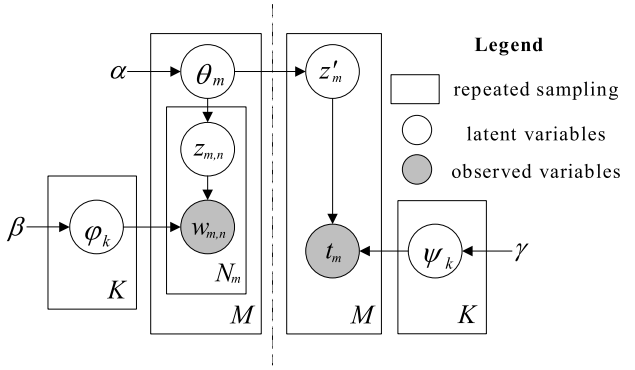

FIGURE 2. Plate notation representing the generative process of the TTM.

- H-3: There exists a multinomial distribution of start time with a Dirichlet prior for each trip purpose.

Based on the hypotheses, the problem is defined as follows. Given that features w and start time t of each group m have been observed and that they are independently generated from K unobservable topics related to trip purpose, it is necessary to estimate the distribution of w and t over each topic and the distribution of topics over each group. **Table 2** lists the notation used for all the related variables.

To solve the problem above, this paper develops a TTM, which is a revision of LDA (the left part in **Fig. 2**). The difference lies in the introduction of the generation process of the start time (the right part in **Fig. 2**). The first reason for this additional consideration is the varied features but identical start time within a group. Thus, the contribution of start time would be concealed by features if they were both measured via the same distribution over the topic. The second is to provide a sufficient basis for trip purpose annotation using the start time distribution in addition to feature distributions. Finally, the ability to predict start time offers an indirect method of model evaluation.

TABLE 3. Algorithm for generating features and start time.

input: $t_m, w_{m,n}, m \in \{1,2,\cdots,M\}, n \in \{1,2,\cdots,N_m\}$
output: $\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\psi}$
for each trip purpose $k \in \{1,2,\ldots,K\}$:
  generate the distribution of features for each trip purpose: $\vec{\varphi}_k \sim \text{Diri}(\beta)$
  generate the distribution of start time for each trip purpose: $\vec{\psi}_k \sim \text{Diri}(\gamma)$
for each group $m \in \{1,2,\ldots,M\}$:
  generate the distribution of trip purpose for each group: $\vec{\theta}_m \sim \text{Diri}(\alpha)$
  for each passenger $n \in \{1,2,\ldots,N_m\}$:
    choose a trip purpose $z_{m,n} = k$: $z_{m,n} \sim \text{Multi}(\vec{\theta}_m)$
    given $z_{m,n} = k$, choose a feature $w_{m,n}$: $w_{m,n} \sim \text{Multi}(\vec{\varphi}_{z_{m,n}=k})$
  choose a trip purpose $z'_m = k$: $z'_m \sim \text{Multi}(\vec{\theta}_m)$
  given $z'_m = k$, choose a start time t_m: $t_m \sim \text{Multi}(\vec{\psi}_{z'_m=k})$

In the plate notation of **Fig. 2**, circles with and without shadows symbolize observed variables and latent variables, respectively, while rectangles represent the repeated random sampling process of each variable from the corresponding probability distribution.

The left part of **Fig. 2** depicts the generation process of the features of each group. First, a multinomial distribution $\text{Multi}(\vec{\theta}_m)$ of trip purposes over group m is determined, with parameters drawn from a Dirichlet distribution $\text{Diri}(\alpha)$. The process is repeated M times. Then, based on the trip purpose $z_{m,n} = k$ sampled from $\text{Multi}(\vec{\theta}_m)$, the feature $w_{m,n}$ is then sampled from $\text{Multi}(\vec{\varphi}_{z_{m,n}=k})$, whose parameters are drawn from $\text{Diri}(\beta)$; this process is repeated N_m times.

The right part is designed to reveal the preference for the start time of each trip purpose. For each group m, based on $z'_m = k$ sampled from $\text{Multi}(\vec{\theta}_m)$, start time t_m is then sampled from $\text{Multi}(\vec{\psi}_{z'_m=k})$, whose parameters are drawn from $\text{Diri}(\gamma)$. Notably, symbols with single quotes (e.g., z'_m, Z' *and* N') differentiate trip purposes generating start time from features. The aforementioned generation processes are summarized in the form of pseudocode in **Table 3**.

B. TRIP PURPOSE INFERENCE BASED ON GIBBS SAMPLING ALGORITHM

Based on Bayesian inference, given the Dirichlet prior parameters α, β and γ, the joint probability distributions of $\mathcal{Z}$, $\mathcal{Z}'$, $\mathcal{W}$ and $\mathcal{T}$ can be formulated as (1) by integrating out $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ according to **Fig. 2** and **Table 3**.

$$p(\mathcal{Z}, \mathcal{Z}', \mathcal{W}, \mathcal{T}|\alpha, \beta, \gamma) = \int_{\boldsymbol{\theta}} p(\mathcal{Z}, \mathcal{Z}'|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\alpha) d\boldsymbol{\theta} \cdot \int_{\boldsymbol{\varphi}} p(\mathcal{W}|\mathcal{Z}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi}|\beta) d\boldsymbol{\varphi} \cdot \int_{\boldsymbol{\psi}} p(\mathcal{T}|\mathcal{Z}', \boldsymbol{\psi}) p(\boldsymbol{\psi}|\gamma) d\boldsymbol{\psi} \quad (1)$$

Since $\mathcal{Z}$ and $\mathcal{Z}'$ are sampled from $\text{Multi}(\boldsymbol{\theta})$, whose prior $\text{Diri}(\alpha)$ is its conjugate distribution, the first integral can be computed as:

$$\begin{aligned} p(\mathcal{Z}, \mathcal{Z}'|\alpha) &= \int_{\boldsymbol{\theta}} p(\mathcal{Z}, \mathcal{Z}'|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\alpha) d\boldsymbol{\theta} \\ &= \prod_{m=1}^{M} \frac{\Delta(\vec{N}_m + \vec{N'}_m + \alpha)}{\Delta(\alpha)} \end{aligned} \quad (2)$$

Similarly, the other integrals are as follows:

$$p(\mathcal{W}|\mathcal{Z},\beta) = \int_{\varphi} p(\mathcal{W}|\mathcal{Z},\varphi)p(\varphi|\beta)d\varphi = \prod_{k=1}^{K}\frac{\Delta(\overrightarrow{N}_k+\beta)}{\Delta(\beta)} \tag{3}$$

$$p(\mathcal{T}|\mathcal{Z}',\gamma) = \int_{\psi} p(\mathcal{T}|\mathcal{Z}',\psi)p(\psi|\gamma)d\psi = \prod_{k=1}^{K}\frac{\Delta(\overrightarrow{N'}_k+\gamma)}{\Delta(\gamma)} \tag{4}$$

where the multi-dimensional beta function $\Delta(\vec{x})$ equals $\frac{\prod_{\vec{x}}\Gamma(x_i)}{\Gamma(\sum_{\vec{x}}x_i)}$ and $\Gamma(x_i)$ is the gamma function.

Using (2)-(4), (1) can be written as below, i.e., the complete-data likelihood function:

$$\begin{aligned} p(\mathcal{Z},\mathcal{Z}',\mathcal{W},\mathcal{T}|\alpha,\beta,\gamma) &== \prod_{m=1}^{M}\frac{\Delta(\overrightarrow{N}_m+\overrightarrow{N'}_m+\alpha)}{\Delta(\alpha)}\prod_{k=1}^{K}\frac{\Delta(\overrightarrow{N}_k+\beta)}{\Delta(\beta)} \\ &\times\prod_{k=1}^{K}\frac{\Delta(\overrightarrow{N'}_k+\gamma)}{\Delta(\gamma)} \end{aligned} \tag{5}$$

As elegant as (5) is, trip purposes $\mathcal{Z}$ and $\mathcal{Z}'$ are difficult to infer exactly using maximum likelihood estimation. Thus, a Markov chain Monte Carlo simulation algorithm, Gibbs sampling, is employed to obtain an approximation. Each $z_{m,n}$ is sampled using the collapsed TTM Gibbs sampler (6), which can be derived from (5) based on Bayes' theorem:

$$\begin{aligned} p(z_{m,n}=k|\mathcal{Z}_{\neg i},\mathcal{Z}',\mathcal{W},\mathcal{T}) &\propto \frac{p(\mathcal{W}|\mathcal{Z})}{p(\mathcal{W}_{\neg i}|\mathcal{Z}_{\neg i})}\frac{p(\mathcal{Z})}{p(\mathcal{Z}_{\neg i})} \\ &= \frac{N^v_{k,\neg i}+\beta}{\sum_{v=1}^{V}N^v_{k,\neg i}+V\beta}\frac{N^k_{m,\neg i}+N'^k_m+\alpha}{\sum_{k=1}^{K}(N^k_{m,\neg i}+N'^k_m)+K\alpha} \\ &\propto \frac{(N^v_{k,\neg i}+\beta)(N^k_{m,\neg i}+1+\alpha)}{\sum_{v=1}^{V}N^v_{k,\neg i}+V\beta} \end{aligned} \tag{6}$$

where i=(m, n) and $\neg i$ denotes dimensions other than i.

In the same way, z'_m can be sampled using the collapsed TTM Gibbs sampler as (7):

$$\begin{aligned} p(z'_m=k|\mathcal{Z},\mathcal{Z}'_{\neg m},\mathcal{W},\mathcal{T}) &\propto \frac{p(\mathcal{T}|\mathcal{Z}')}{p(\mathcal{T}_{\neg m}|\mathcal{Z}'_{\neg m})}\frac{p(\mathcal{Z},\mathcal{Z}')}{p(\mathcal{Z},\mathcal{Z}'_{\neg m})} \\ &= \frac{N'^l_{k,\neg m}+\gamma}{\sum_{l=1}^{L}N'^l_{k,\neg m}+L\gamma}\frac{N^k_m-1+\alpha}{\sum_{k=1}^{K}(N^k_m-1)+K\alpha} \\ &\propto \frac{(N'^l_{k,\neg m}+\gamma)(N^k_m-1+\alpha)}{\sum_{l=1}^{L}N'^l_{k,\neg m}+L\gamma} \end{aligned} \tag{7}$$

The posterior probabilities of $z_{m,n}$ and z'_m converge to a stationary distribution after the burn-in period; then, the counters N^v_k, N'^l_k, N^k_m and N'^k_m can be used to estimate $\vec{\theta}_m$, $\vec{\varphi}_k$ and $\vec{\psi}_k$, which follow a Dirichlet distribution according to (2)-(4):

$$\hat{\theta}^k_m = \frac{N^k_m+N'^k_m+\alpha}{\sum_{k=1}^{K}(N^k_m+N'^k_m)+K\alpha} \tag{8}$$

$$\hat{\varphi}^v_k = \frac{N^v_k+\beta}{\sum_{v=1}^{V}N^v_k+V\beta} \tag{9}$$

$$\hat{\psi}^l_k = \frac{N'^l_k+\gamma}{\sum_{l=1}^{L}N'^l_k+L\gamma} \tag{10}$$

The TTM can predict start time as well, which means that given a new group, denoted as a vector $\vec{w}_{\tilde{m}}$, a query of probability distribution $t_{\tilde{m}}$ can be calculated by sampling the trip purpose distribution $\vec{\theta}_{\tilde{m}}$ for $\vec{w}_{\tilde{m}}$ beforehand:

$$\begin{aligned} p(t_{\tilde{m}}=l|\overrightarrow{w}_{\tilde{m}}) &\propto \sum_{k=1}^{K}p(t_{\tilde{m}}=l|z'_{\tilde{m}}=k)p(z'_{\tilde{m}}=k|\overrightarrow{\theta}_{\tilde{m}}) \\ &= \sum_{k=1}^{K}(\hat{\psi}^l_k\cdot\theta^k_{\tilde{m}}) \end{aligned} \tag{11}$$

The source code of the TTM is available from *https://github.com/jianpeiqian/TripPurposeInference*.

C. DESIGN OF FEATURES AND RECONSTRUCTION OF GROUPS

The start time is discretized preliminarily, including five non-overlapping time slots in total (*Spring-Festival (SpringF)*, *Holiday*, *Summer*, *Weekend* and *Weekday*).

Regarding the feature design, the principles learned from *text mining* should be clarified beforehand:

- The features should be defined as discrete variables;
- The values of each feature should have semantic meanings related to trip purposes;
- The whole set of features should not be too small to cause the "*short-text dilemma*";
- The whole set of values, i.e., "vocabulary", should not be too large to cause the "*low co-occurrence*".

Thus, **Table 4** lists candidate features extracted from trip records based on ticket sales data. Generally, the features are

TABLE 4. List of alternative features of a passenger.

Aspects	Feature	Description
Demographic features	AgeGender	Dualistic variable combining age bracket and gender
Experience features	Duration	How long is the first travel to current destination
	Frequency	Average travel times to current destination
	MinInterval	Minimum interval of two successive trip to current destination
	TimeHabit	Whether the current destination has been visited at a certain start time more than once
	LastTime	Last start time of visiting current destination
Co-travel network features	Degree	Total companions when visiting current destination

TABLE 5. List of the values of each feature.

Value	Description
Minor	Age <18
Teen	$18 \leq$ Age ≤ 25
Youth	$26 \leq$ Age ≤ 35
Prime	$36 \leq$ Age ≤ 45
Middle	$46 \leq$ Age ≤ 55
Senior	Age >55
Male	Gender is male
Female	Gender is female
Never	Duration $=0$
WithinOneYear	$0<$ Duration <400 (days)
Ealier	Duration ≥ 400 (days)
Freq $<$ *OneQuarter*	Frequency $>1/110$ (/days)
Freq $>$ *OneYear*	Frequency $<1/400$ (/days)
MinOneWeek	MinInterval ≤ 7 (days)
MinOneMonth	$8 \leq$ MinInterval ≤ 30 (days)
MinOneYear	$345 \leq$ MinInterval ≤ 390 (days)
FreqSpringFest	TimeHabit of *Spring-Festival* is true
FreqSummer	TimeHabit of *Summer* is true
FreqHoliday	TimeHabit of *Holiday* is true
FreqWeekday	TimeHabit of *Weekday* is true
FreqWeekend	TimeHabit of *Weekend* is true
LastSpringFest	LastTime $=$ *Spring-Festival*
LastSummer	LastTime $=$ *Summer*
LastHoliday	LastTime $=$ *Holiday*
LastWeekday	LastTime $=$ *Weekday*
LastWeekend	LastTime $=$ *Weekend*
Couple	Degree $=2$
Crowds	Degree >4

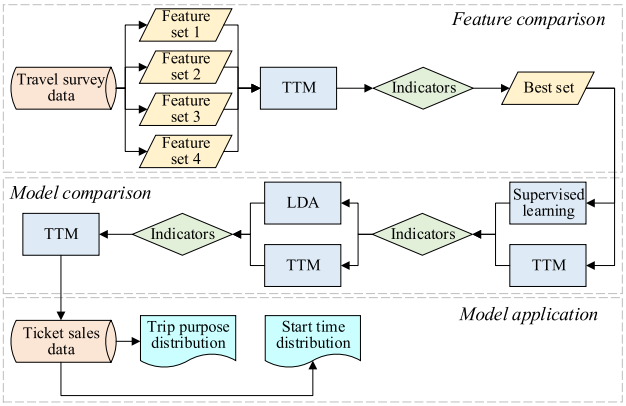

FIGURE 3. Framework of the experiment based on the travel survey.

TABLE 6. Samples and proportions regarding different trip purposes of groups of different sizes.

Group size	Work/ study	Official	Personal	Journey	Total
2	140 (42.4%)	39 (11.8%)	51 (15.5%)	100 (30.3%)	330 (100%)
3, 4, 5	39 (24.4%)	19 (11.9%)	23 (14.4%)	79 (49.4%)	160 (100%)
6, 6+	7 (14%)	8 (16%)	2 (4%)	33 (66%)	50 (100%)

classified into three types: demographic features are chosen to reveal the social relationships of the members based on their age and gender [8]; experience features may indicate whether a passenger is habitual from the perspective of time and space [7]; and co-travel network features reflect the potential number of companions. In addition, features have no orders in a group since the TTM is a *bag-of-words* model.

Table 5 lists the values and provides descriptions of each feature. For example, for a group with two members described as {{'*YouthMale*', '*Couple*', '*FreqWeekday*', '*LastWeekday*', '*WithinOneYear*', '*MinOneMonth*', }, {'*PrimeMale*', '*Couple*', '*LastWeekday*', '*WithinOneYear*'}}, a plausible guess is that they were involved in official business travel for the past year. Clearly, more creditable conclusions need to be drawn with consideration of the co-occurrence of the features in a group using the TTM.

Reconstruction of the group is the fundamental task of the TTM; nevertheless, the group remains to be identified from ticket sales data since trip records are initially organized by tickets. Motivated by the desire to sit side by side, passengers of the same group often purchase tickets in a single transaction, leading to the precisely same timestamp for each ticket. Therefore, the combination of the *deal time* and *schedule bus ID* fields is chosen as the primary key to reconstruct groups with consideration of accuracy and efficiency.

IV. EXPERIMENTS

A. WEB-BASED TRAVEL SURVEYS

Before applying the TTM to ticket sales data, the explanatory power of features and the accuracy of the TTM compared to that of traditional methods must be evaluated. Therefore, two web-based travel surveys on inter-city trips via road passenger transport in China were conducted in January 2018 and January 2020. Based on the travel survey data with ground truth, two parallel experiments, i.e., feature comparison and model comparison, are illustrated in **Fig. 3**.

The earlier travel survey revealed that 55.64% of respondents reported having at least one companion. The latter focused on trips in groups, and 540 samples were obtained for the subsequent experiments. Trip purpose is segmented as 'return home after work/study or go to work/study' (work/study), 'official business' (official), 'personal business' (personal), and 'journey'. According to **Table 6**, 'work/study' and 'journey' account for 73.7% of travel by road passenger transport, whereas 'official' and 'personal' travel is relatively rare. In addition, the proportions of trip purposes differ among groups of different sizes, which implies the significance of co-travel network features. The remaining items are designed in accordance with the features that can be obtained from ticket sales data, as shown in **Table 4**.

B. FEATURE COMPARISONS

Although the features listed in **Table 4** were investigated in the travel survey, not all of them can be obtained from ticket sales data (i.e., demographic features) or be instantaneously calculated (i.e., co-travel network features). Therefore, to evaluate the explanatory power when the additional features are absent, the features are divided into four sets:

- *set 1*: The full feature set with all candidate features included;

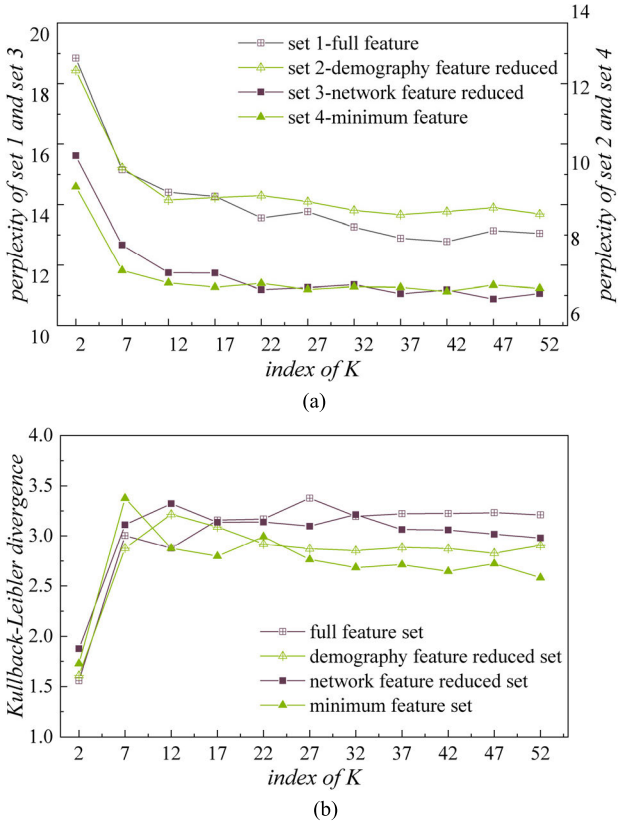

FIGURE 4. (a) Comparison of *perplexity* for different sets. (b) Comparison of KL_{avg} for different sets.

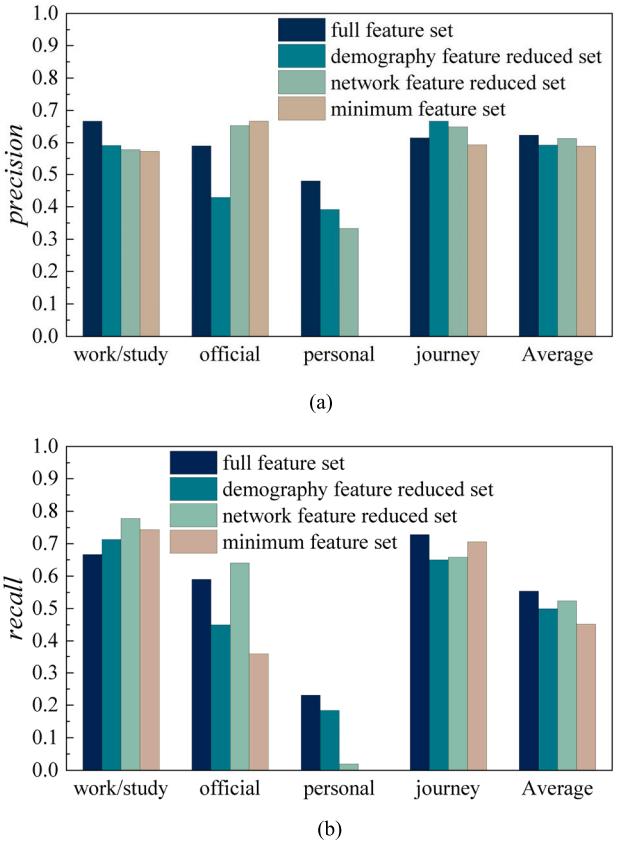

FIGURE 5. (a) Comparison of *precision* for different sets. (b) Comparison of *recall* for different sets.

- *set 2*: The full feature set with demographic features reduced;
- *set 3*: The full feature set with co-travel network features reduced;
- and *set 4*: The minimum feature set with only experience features retained.

Two indicators are considered to identify the best feature set. First, *perplexity* and *Kullback-Leibler divergence* (*KL*) are introduced to evaluate the robustness of the topic models (TTM and LDA), given a series of numbers of hyperparameters K. *Perplexity* is defined as the reciprocal of the geometric mean of the likelihood of each feature using the training set D_{test}, and a lower *perplexity* score indicates a higher certainty of trip purpose with regard to each group. The essence of *KL* is relative entropy, which is used to measure the difference in the probability distribution. A higher value of KL_{avg} indicates that trip purpose has a higher dissimilarity in terms of *KL* on average. The *perplexity* and KL_{avg} are calculated according to (12) and (13), respectively.

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{\tilde{m}=1}^{M}\log p(\vec{w}_{\tilde{m}})}{\sum_{\tilde{m}=1}^{M} N_{\tilde{m}}}\right\} \tag{12}$$

$$KL_{avg} = \left(\sum_{i=1}^{K}\sum_{j=1}^{K}(\sum_{v=1}^{V}\hat{\varphi}_i^v \log\frac{\hat{\varphi}_i^v}{\hat{\varphi}_j^v})\right)/K^2 \tag{13}$$

Overall, set 1 outperforms the others to some extent. On the one hand, **Fig. 4(a)** shows that a larger K yields a better *perplexity* for each feature set, but the greatest drop in *perplexity* is observed for *set 1*. Thus, the TTM will generate fine-grained clusters if all the features are considered, and a larger K is preferable. On the other hand, **Fig. 4(b)** shows that the TTM becomes more robust as K grows for *set 1* since KL_{avg} reaches a maximum at K=*27* and then remains stable. However, for the reduced sets, *set 4* in particular, the clusters become similar, as observed from the decrease in KL_{avg} a K grows. Therefore, a larger K will reduce the generalization ability, and the best K should be determined by balancing *perplexity* and KL_{avg} if the full feature set is inaccessible.

The second set of indicators for feature selection includes *precision* and *recall* and represents the classification power when different features are considered.

In summary, more evidence that support *set 1* is better is obtained. Nevertheless, set 3 is acceptable if the job is restricted by data collection or computational resource. **Figure 5(a)** indicates that the average *precision* of *set 1* is 62.3%, whereas if demographic features or co-travel network features (or both) are reduced, the precision decreases to 59.2%, 61.3% and 58.7%, respectively. Specifically, 'personal' cannot be well predicted based on *set 1*; the *precision* of 'official' decreases substantially if demographic features are reduced; and the *precision* of 'work/study' and

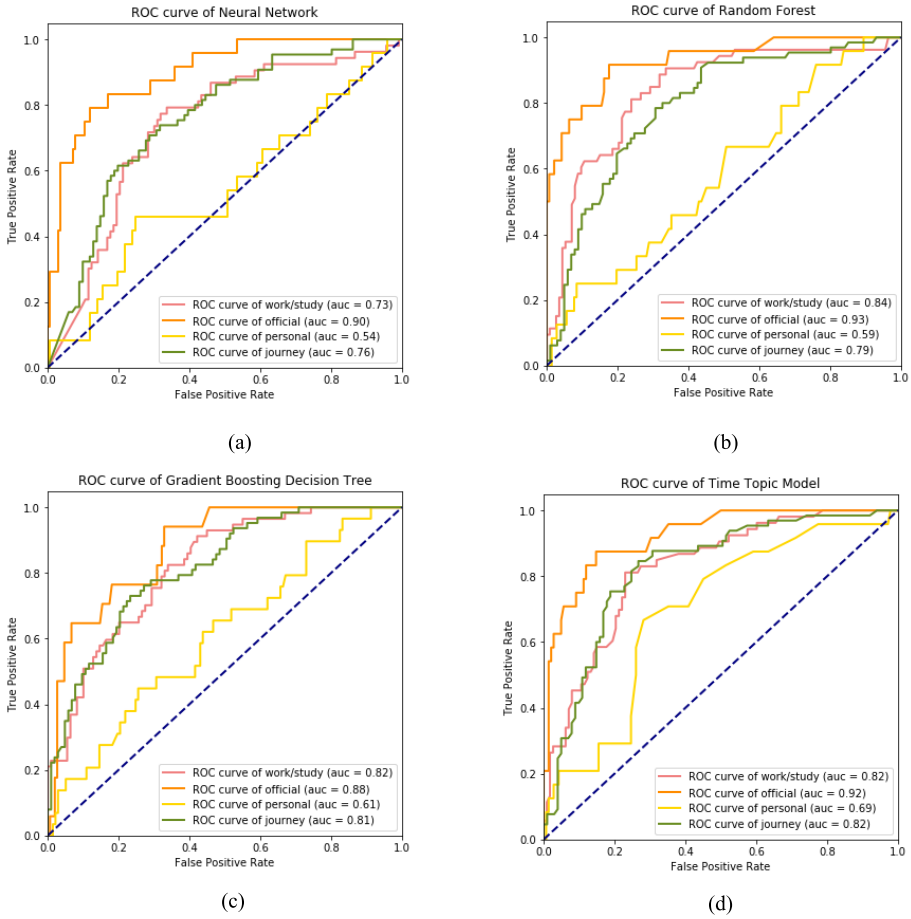

FIGURE 6. (a) ROC curve of the neural network model. (b) ROC curve of the random forest model. (c) ROC curve of the gradient boosting decision tree model. (d) ROC curve of the time topic model.

'journeys' is relatively stable. **Figure 5(b)** indicates that, similar to *precision*, the average *recall* of *sets 1* and *3* is better than that of *sets 2* and *4*. In fact, a significant difference in *recall* is observed because an excessive number of samples are labelled 'work/study' or 'journey' due to imbalanced sampling.

On the basis of the above analysis, all the features, i.e., *set 1*, will be considered in the following discussions.

C. MODEL COMPARISONS

1) TTM AND SUPERVISED LEARNING METHODS

TTM utilizes features from the perspective of the group, which makes it different from supervised learning, which considers individual features. To compare these two methods, the receiver operating characteristic curve (ROC) curve, which is drawn based on the *true positive rate* (TPR) and *false positive rate* (FPR), is adopted. Moreover, the area under the ROC curve (AUC) is considered since it combines *precision* and *recall* and is insensitive to imbalanced sampling. Specifically, a more generalizable method will have an ROC curve located much closer to the top left of the graph, even for a low threshold. In contrast, a curve closer to the dashed diagonal is almost equivalent to random guessing.

On the basis of the summary in **Table 1**, ANN and two decision tree ensemble methods, i.e., RF and gradient boosting decision tree (GBDT), are chosen as the baseline models. The discrete features are transformed using one-hot encoding beforehand. Then, hold-out cross-validation with a 70%/30% split is applied.

Figure 6(a)-(d) demonstrates that the TTM is no worse than the baselines, despite the limited samples; moreover, the TTM obtains a more balanced prediction for each trip purpose. Notably, the AUC of 'personal' increases from 0.61 to 0.69 if the TTM is employed instead of GBDT. By contrast, neither of the other two methods can recognize 'personal', as the ROC curves follow the diagonal. Meanwhile, the ability to predict 'work/study' and 'journey' is similar, and the TTM, RF and GBDT are superior to the ANN. Additionally, 'official' can always be well predicted, as the AUC ranges from 0.88 to 0.93.

2) TTM AND LDA

According to hypothesis H-3 proposed in Section III, the TTM enriches LDA by utilizing not only feature distribution $\boldsymbol{\varphi}$ but also start time distribution $\boldsymbol{\psi}$ to annotate trip purpose. The efficiency of this additional consideration can

TABLE 7. Data descriptions.

Year	2014	2015	2016	2017	2018.1-3	Total
The original data						
Number of trips	12,335,878	10,933,968	9,053,552	7,934,290	1,467,152	41,728,470
Number of individuals	7,129,759	6,391,426	5,375,714	4,852,757	1,243,989	16,918,470
Number of groups (with 2 members or more)	2,127,221	1,814,942	1,359,684	1,126,785	188,941	6,617,573
Percentage of trips performed in groups	42.15%	39.67%	35.56%	33.75%	30.52%	38.06%
The study sample of passengers travelling to Shanxi Province, China						
Number of trips	474,396	454,166	462,256	458,224	95,199	1,944,241
Number of individuals	348,249	332,088	331,573	326,013	88,468	1,047,520
Number of groups (with 2 members or more)	98,908	85,544	69,361	58,439	8,222	320,474
Percentage of the trips performed in groups	49.32%	44.08%	34.46%	29.28%	19.40%	38.38%

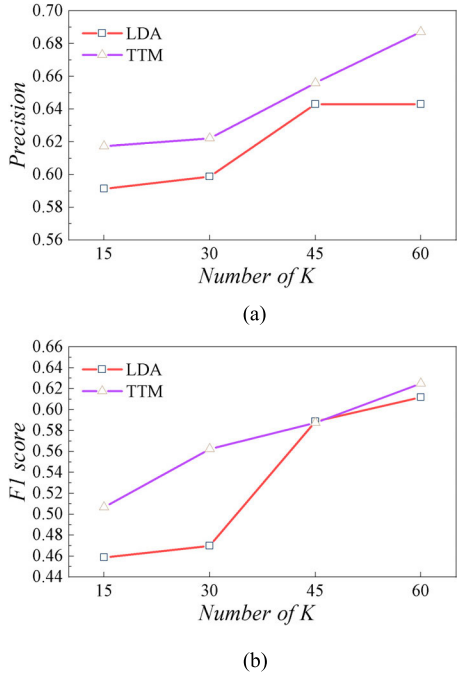

FIGURE 7. (a) *Precision* of the TTM and LDA. (b) *F1 score* of the TTM and LDA.

be discussed in terms of the algorithm complexity: for both the TTM and LDA, the complexity is $O(K * N *$ Iterations), using Gibbs sampling. Thus, the classification performance is improved. **Figure 7(a)** shows that *precision* increases with increasing K, and the *precision* of the TTM is greater than that of LDA for the entire set of K. In addition, the $F1$ score of LDA is almost the same as that of TTM when K becomes larger, but it is worse for a smaller K, as shown in **Fig. 7(b)**.

V. APPLICATION

A. LARGE-SCALE TICKET SALES DATA

Large-scale ticket sales data were obtained from the Beijing Road Passenger Transport System (BJRPTS), which manages more than 900 state-wide bus routes carrying approximately ten million passengers annually. Since 2014, in Beijing, passengers have been required to provide their personal information, such as their identity card, when purchasing tickets, which offers us a chance to trace historical trips. After the elimination, anonymization and cleaning process, only a masking identifier and demographic features (age and gender) of a passenger are retained for this study, in addition to ticket information.

The original data include 16,918,470 individuals who had taken 41,724,840 trips from January 1, 2014, to March 31, 2018. The original data accounted for 95.75% of the total records. General descriptions of the data are provided in **Table 7**. A total of 38.06% of the trips are performed in groups.

A sample of passengers travelling to Shanxi Province, China, is selected for the following discussions, considering that road passenger transport played an important role before 2018 when only one high-speed rail route was available. Only the groups with 2 members or more are considered in this study, and nearly half of the trip purposes in 2014 can be inferred from the TTM, as shown in **Table 7**.

Two power laws are observed based on the original data, as plotted in double logarithmic coordinates. **Figure 8(a)** depicts the distribution of total travel times of the 16,918,470 individuals. Those with only one trip account for 60.27% of the data. However, the remaining passengers, defined as frequent passengers, account for 75.56% of all trips. **Fig. 8(b)** depicts the distribution of the total visiting cities of individuals. A total of 79.70% of passengers visit one city, which is also helpful to explore the experience features.

B. MODEL ESTIMATION

The TTM is estimated on the Dell® workstation (CPU: Intel® Xeon® E5-2640 @2.50 GHz; RAM: 32 GB). This paper chooses 10% of the samples of Shanxi to evaluate the *perplexity* and *Jensen-Shannon divergence* (*JS*) in the condition of K ranging from 2 to 100. *JS* measures the dissimilarity of two probability distributions on a scale of 0 to 1 according to (14) and is a modified indicator of *KL*. The higher the *JS* is, the higher the dissimilarity indicates.

$$JS(P_1||P_2)=\frac{1}{2}KL(P_1||\frac{P_1+P_2}{2})+\frac{1}{2}KL(P_2||\frac{P_1+P_2}{2}) \quad (14)$$

where P_1, P_2 denote any two probability distributions.

In **Fig. 9,** there is no remarkable decrease in *perplexity* when K is greater than 50: *JS* fluctuates around approximately 0.47. By weighing the indicators and the generalization performance of the TTM, this paper selects $K = 50$ for further study.

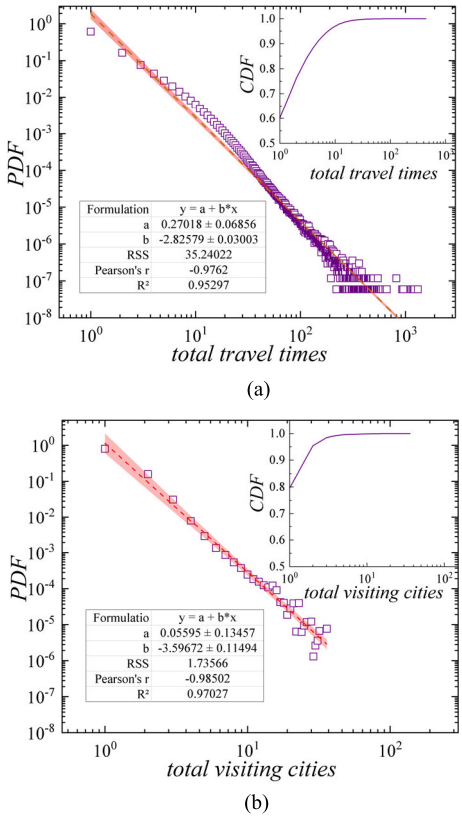

FIGURE 8. **(a) Distribution of the total travel times of individuals. (b) Distribution of the total visiting cities.**

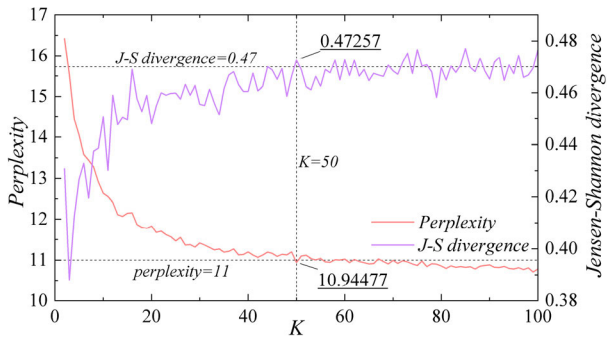

FIGURE 9. ***Perplexity*** **and** ***JS*** **under the condition of different K.**

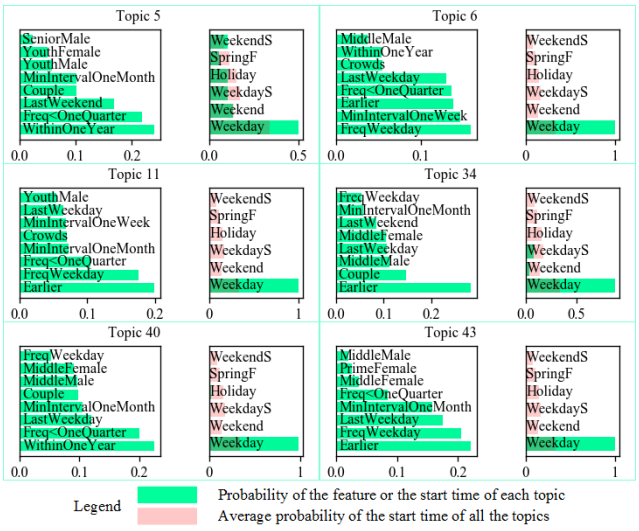

FIGURE 10. **The features and start time distribution of official groups.**

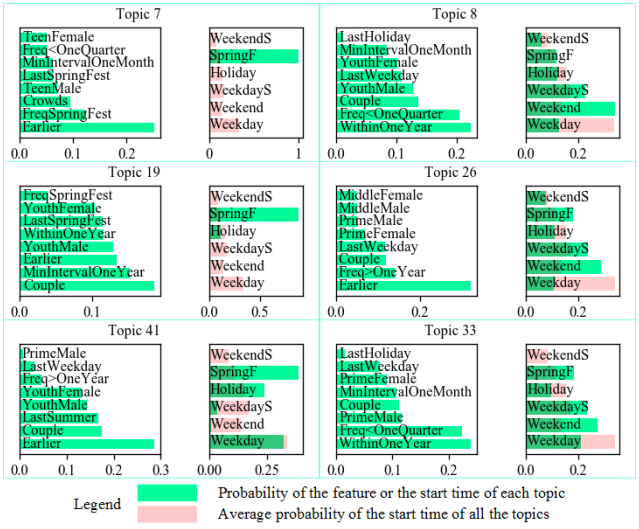

FIGURE 11. **The features and start time distribution of typical work groups.**

C. ANNOTATING TRIP PURPOSE

By combining feature distribution $\boldsymbol{\varphi}$ and start time distribution $\boldsymbol{\psi}$ for the 50 topics, each is annotated with a trip purpose. In summary, 46 topics are clustered into four primary or eight subtypes of trip purpose, in accordance with the categorization predefined in the web-based travel survey. In addition, the remaining topics beyond existing knowledge are detected as anomalies. Notably, the four primary types, i.e., 'work/study', 'official', 'personal' and 'journey' accounted for 31.89%, 12.38%, 18.86% and 29.57% of all trips, similar to the results of the web-based travel survey.

In the left part of **Fig. 10-18**, horizontal bars in green represent the top eight features with a higher probability of $\boldsymbol{\varphi}$. The right part visualizes the probability of $\boldsymbol{\psi}$ and the average probability overall, indicated as green and pink bars, respectively. Notably the start time *Summer* is separated into *weekday in summer* (*WeekdayS*) and *weekend in summer* (*WeekendS*).

Cluster 1 (Official Business (Official)): It is already illustrated in **Fig. 6** that the 'official' travel has evident features that are easy to identify. Therefore, this paper first annotates the topics with a $\vec{\psi}_k$ dominated by '*Weekday*' as 'official'. Meanwhile, a higher probability than the average of the other start time and the occurrence of '*Minor*', '*Teen*' or '*Senior*' passengers are not considered. As a result, six topics are annotated, accounting for 12.38% of the groups in the test set on the condition that the topic taking the highest probability of θ_m^k is regarded as the trip purpose of group m. **Fig. 10** shows that this subtype generally shares the features '*Freq* $<$ *OneQuarter*', '*FreqWeekday*', '*Earlier*' and '*MinInterval-OneMonth*'.

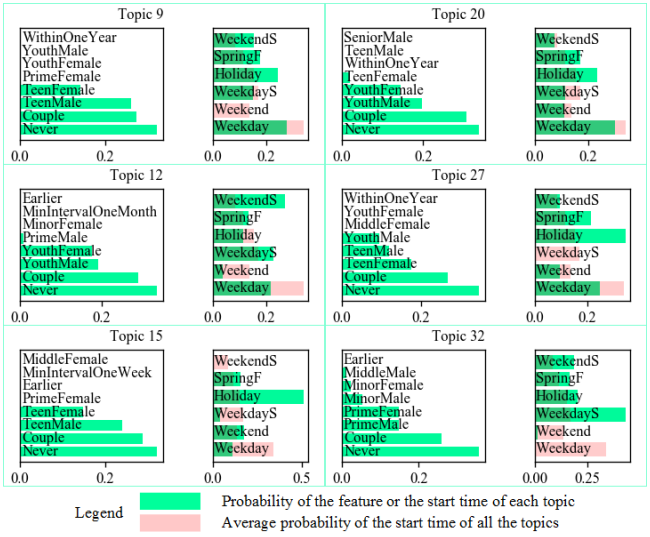

FIGURE 12. The features and start time distribution of atypical work groups.

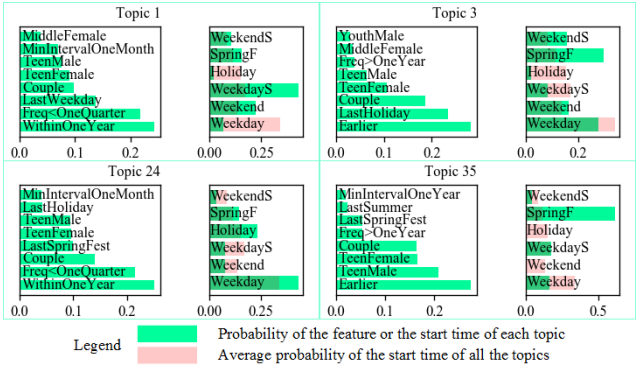

FIGURE 13. The features and start time distribution of the study groups.

Cluster 2 (Returning Home of Typical Migrant Workers (Typical Work)): Then, the topics related to returning home after study or work are annotated. Migrant workers who pursue opportunities in Beijing seasonally return to their hometown like migrating birds. Topics 7 and 19 are the most typical cases in which passengers have the features '*Earlier*' and '*FreqSpringFest*'. Generally, if '*Earlier*' or '*WithinOneYear*' is present and the probability of '*SpringF*' is greater than average, the topics are classified as typical work. The six topics shown in **Fig. 11** account for 10.95% of the groups.

Cluster 3 (Returning Home of Migrant Workers Mixed With Journey (Atypical Work)): It is imaginable that not all migrant workers always take the bus, especially in the case of travelling in groups, when a private car is a more comfortable choice. Therefore, if both '*SpringF*' and '*Holiday*' are higher than average and '*Weekday*' is lower than average, a trip could be regarded as atypical work. This mixture type may include journey travel, as shown in **Fig. 12**.

Cluster 4 (Returning Home of Undergraduates (Study)): Beijing has more than one million undergraduates who regularly return home during winter and summer vacation and occasionally during other holidays. In this sense, topics in

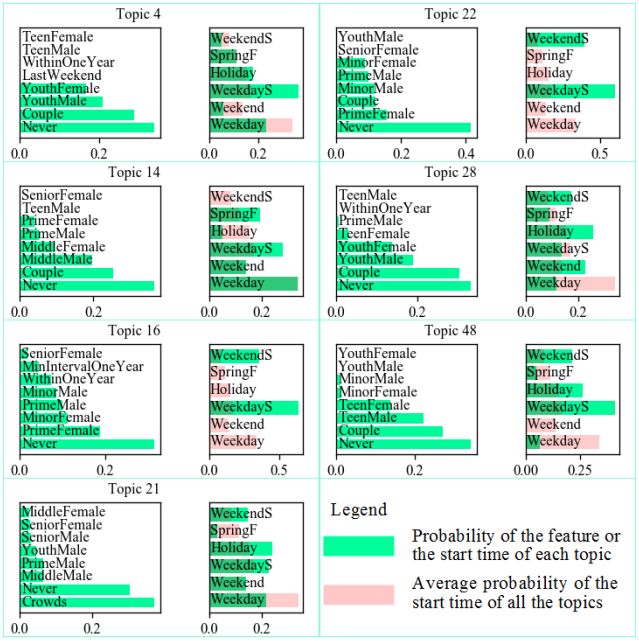

FIGURE 14. The features and start time distribution of typical journey groups.

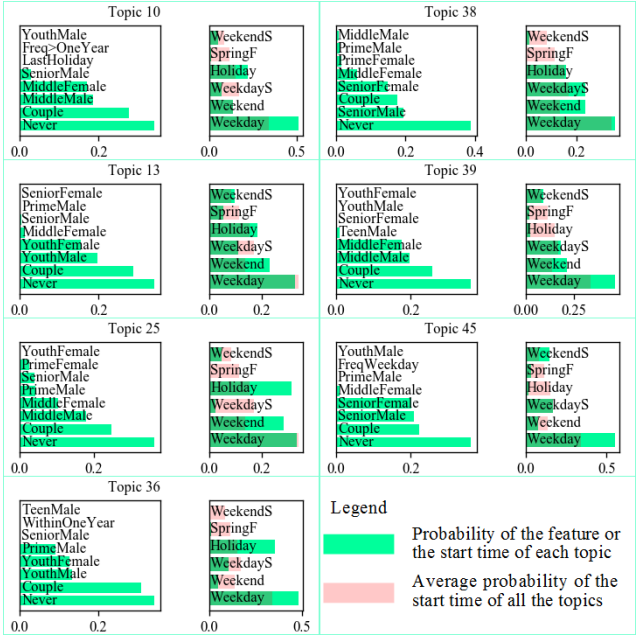

FIGURE 15. The features and start time distribution of atypical journey groups.

Fig. 13 can be annotated as 'study', accounting for 7.27% of all the groups. The $\vec{\varphi}_k$ of these topics shows a preference for '*Earlier*' or '*WithinOneYear*' for '*Teen*' passengers, meanwhile '*SpringF*' has a higher probability than average. In particular, topic 3 and topic 35 represent typical 'study' travel.

Cluster 5 (Typical Journey): There are five legal holidays and a summer vacation during the traditional tourist season from April to October, when many visitors are produced or attracted. Therefore, seven topics in **Fig. 14**, accounting

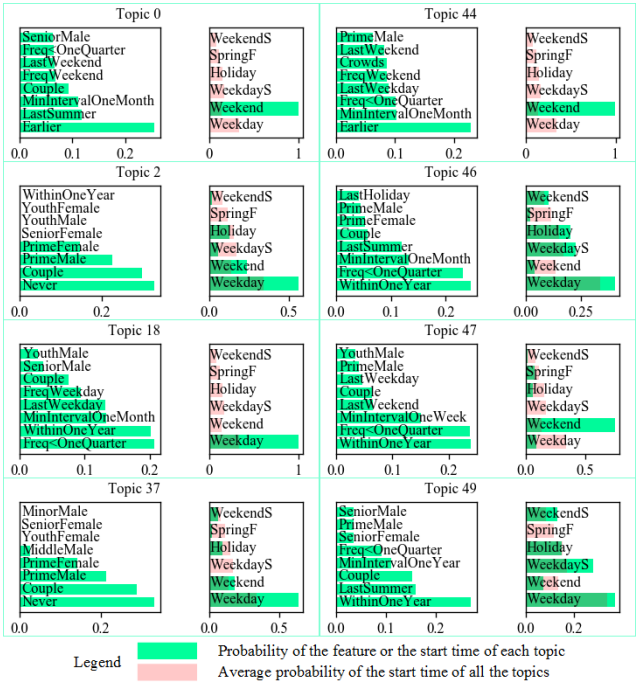

FIGURE 16. The features and start time distribution of personal groups.

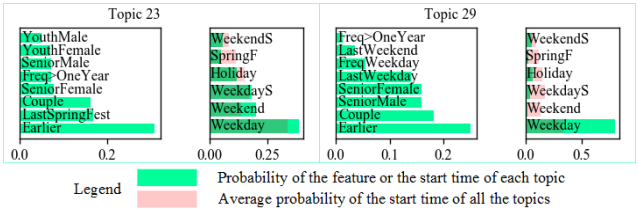

FIGURE 17. The features and start time distribution of social groups.

for 14.35% of the groups, are annotated as typical journeys for passengers who travel on a '*Holiday*', '*WeekdayS*' or '*WeekendS*' most of the time and have '*Never*' gone to the destination before. For example, topics 22 and 48 describe '*Teen*' tourists accompanied by their parents or friends, perhaps during summer vacation.

Cluster 6 (Atypical Journey Mixed With Personal Business (Atypical Journey)): However, some tourists prefer weekdays to holidays to avoid congestion, especially for those who have more flexible schedules, such as 'Senior' in topics 38 and 45. '*Never*' is still a necessary condition for judging a journey. In addition, the topics with the greatest occurrence on a '*Weekday*' and a higher probability of '*Holiday*' than average are considered, as shown in **Fig. 15**. *Atypical* is used to depict this type of journey because it is inevitably mixed with personal business.

Cluster 7 (Personal Business (Personal)): In Section IV, personal business is shown to be the most difficult type of trip purpose to infer. In the TTM, 'business', 'work/study' and 'journey' are annotated sequentially, and the remaining topics are classified as 'personal business'. **Fig. 16** reveals a group composed of '*Prime*' passengers sharing the feature '*Earlier*' or '*WithinOneYear*' travel at '*Weekday*' or '*Weekend*'.

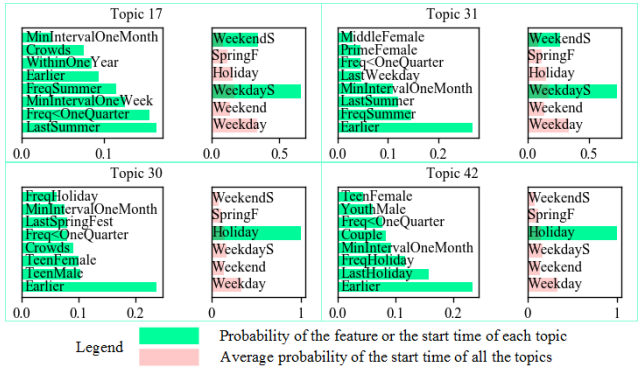

FIGURE 18. The features and start time distribution of anomaly groups.

Personal business reflects travel that is frequent but irregular, accounting for 15.43% of the groups.

Cluster 8 (Visiting Relatives (Social)): In addition to the broad and vague classification of personal business, two particular topics, i.e., topics 23 and 29 in **Fig. 17,** are detected, with '*Senior*' '*Couple*' frequently travelling on a '*Weekday*'. This type of trip could involve visiting relatives, accounting for 3.43% of the data.

Cluster 9 (Anomaly): In addition to the four primary types, reflecting the general understanding of the patterns in inter-city trips, two anomalies are detected using the TTM, as shown in **Fig. 18**. The first is composed of topics 17 and 31, which depict a crowd of passengers frequently travelling in the 'summer'. The second is composed of topics 30 and 42, which depict groups who have frequently travelled on a '*Holiday*' since '*Earlier*'. Though the 'anomaly' type cannot be explained straightforwardly, it should not be ignored in the operation and management of road passenger transport because it accounts for up to 7.31%.

VI. CONCLUSION

In this paper, we develop the TTM for inter-city trip purpose inference. On the one hand, the TTM postulates that trip purpose can be inferred from the co-occurrence of features in a group, including demographic, experience and co-travel network features extracted from ticket sales data. On the other hand, the TTM modifies LDA by incorporating generation process of start time. Three multinomial distributions, as proposed in $H1$-$H3$, are estimated using a Gibbs sampler.

Before applying the TTM to ticket sales data, comparison experiments are conducted based on web-based travel surveys. First, dividing the features into four sets proves that the TTM is more robust when the full feature set is considered. However, when resources are limited, co-travel network features can be reduced, and the best K should be determined by balancing *perplexity* and KL_{avg}. Second, compared to three baseline supervised learning methods, the TTM showed balanced predictive power: not only trip purposes with typical patterns but also atypical patterns or anomalies can be inferred. By comparison, the baselines are incapable of recognizing personal business, and the AUC varies greatly

for different trip purposes. Third, with the extra information provided by the start time distribution, TTM improves the performance of LDA under conditions of different K, and the estimated efficiency is not inferior to that of LDA.

The TTM is applied based on large-scale ticket sales data obtained from the road passenger transport system, the market share of which has been shrinking recently in China. Using 320,487 samples from Beijing to Shanxi Province as a case study, two kinds of anomalies, in addition to four primary trip purposes, are detected, and the proportion of trip purposes in the test set is in accordance with the web-based travel survey. Through the penetration of trip purposes of passengers, this paper makes it possible for relevant operators to offer further personalized services.

However, some limitations remain. The TTM is applied only to groups with 2 members or more: the ability to infer trip purpose from all the records is unclear. For future research, the concept of the group could be extended to consider the potential co-travel network rather than just the current group.

REFERENCES

[1] L. Shen and P. R. Stopher, "Review of GPS travel survey and GPS data-processing methods," *Transp. Rev.*, vol. 34, no. 3, pp. 316–334, May 2014, doi: 10.1080/01441647.2014.903530.

[2] L. Gong, X. Liu, L. Wu, and Y. Liu, "Inferring trip purposes and uncovering travel patterns from taxi trajectory data," *Cartography Geographic Inf. Sci.*, vol. 43, no. 2, pp. 103–114, Mar. 2016, doi: 10.1080/15230406.2015.1014424.

[3] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 240–250, Sep. 2015, doi: 10.1016/j.trc.2015.02.018.

[4] A. Ermagun, Y. Fan, J. Wolfson, G. Adomavicius, and K. Das, "Real-time trip purpose prediction using online location-based search and discovery services," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 96–112, Apr. 2017, doi: 10.1016/j.trc.2017.01.020.

[5] Y. Cui, C. Meng, Q. He, and J. Gao, "Forecasting current and next trip purpose with social media data and Google places," *Transp. Res. C, Emerg. Technol.*, vol. 97, pp. 159–174, Dec. 2018, doi: 10.1016/j.trc.2018.10.017.

[6] F. Wang, J. Wang, J. Cao, C. Chen, and X. Ban, "Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example," *Transp. Res. C, Emerg. Technol.*, vol. 105, pp. 183–202, Aug. 2019, doi: 10.1016/j.trc.2019.05.028.

[7] A. Alsger, A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman, "Public transport trip purpose inference using smart card fare data," *Transp. Res. C, Emerg. Technol.*, vol. 87, pp. 123–137, Feb. 2018, doi: 10.1016/j.trc.2017.12.016.

[8] Y. Lin, H. Wan, R. Jiang, Z. Wu, and X. Jia, "Inferring the travel purposes of passenger groups for better understanding of passengers," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 235–243, Feb. 2015, doi: 10.1109/TITS.2014.2329422.

[9] J. Wolf, R. Guensler, and W. Bachman, "Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1768, no. 1, pp. 125–134, Jan. 2001, doi: 10.3141/1768-15.

[10] L. Shen and P. R. Stopher, "A process for trip purpose imputation from global positioning system data," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 261–267, Nov. 2013, doi: 10.1016/j.trc.2013.09.004.

[11] T. Feng and H. J. P. Timmermans, "Extracting activity-travel diaries from GPS data: Towards integrated semi-automatic imputation," *Procedia Environ. Sci.*, vol. 22, pp. 178–185, Jan. 2014, doi: 10.1016/j.proenv.2014.11.018.

[12] 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," *Transp. Res. C, Emerg. Technol.*, vol. 68, pp. 285–299, Jul. 2016, doi: 10.1016/j.trc.2016.04.005.

[13] D. M. Blei, A. Y. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vols. 4–5, no. 3, pp. 993–1022, 2003, doi: 10.1162/jmlr.2003.3.4-5.993.

[14] S. Reumers, F. Liu, D. Janssens, M. Cools, and G. Wets, "Semantic annotation of global positioning system traces," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2383, no. 1, pp. 35–43, Jan. 2013, doi: 10.3141/2383-05.

[15] S. G. Lee and M. Hickman, "Trip purpose inference using automated fare collection data," *Public Transp.*, vol. 6, nos. 1–2, pp. 1–20, Apr. 2014, doi: 10.1007/s12469-013-0077-5.

[16] M. G. S. Oliveira, P. Vovsha, J. Wolf, and M. Mitchell, "Evaluation of two methods for identifying trip purpose in GPS-based household travel surveys," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2405, no. 1, pp. 33–41, Jan. 2014, doi: 10.3141/2405-05.

[17] Y. Kim, F. C. Pereira, F. Zhao, A. Ghorpade, P. C. Zegras, and M. Ben-Akiva, "Activity recognition for a smartphone based travel survey based on cross-user history data," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 432–437.

[18] M. Allahviranloo and W. Recker, "Mining activity pattern trajectories and allocating activities in the network," *Transportation*, vol. 42, no. 4, pp. 561–579, Jul. 2015, doi: 10.1007/s11116-015-9602-5.

[19] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transp. Res. C, Emerg. Technol.*, vol. 46, pp. 179–191, Sep. 2014, doi: 10.1016/j.trc.2014.05.012.

[20] L. Montini, N. Rieser-Schüssler, A. Horni, and K. W. Axhausen, "Trip purpose identification from GPS tracks," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2405, no. 1, pp. 16–23, Jan. 2014, doi: 10.3141/2405-03.

[21] M. Janzen, M. Vanhoof, K. W. Axhausen, and Z. Smoreda, "Purpose imputation for long-distance tours without personal information," in *Proc. 96th Annu. Meeting Transp. Res. Board*, Washington, DC, USA, 2017, pp. 2417–2426.

[22] G. Xiao, Z. Juan, and C. Zhang, "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization," *Transp. Res. C, Emerg. Technol.*, vol. 71, pp. 447–463, Oct. 2016, doi: 10.1016/j.trc.2016.08.008.

[23] W. Ectors, S. Reumers, W. D. Lee, K. Choi, B. Kochan, D. Janssens, T. Bellemans, and G. Wets, "Developing an optimised activity type annotation method based on classification accuracy and entropy indices," *Transportmetrica A, Transp. Sci.*, vol. 13, no. 8, pp. 742–766, Sep. 2017, doi: 10.1080/23249935.2017.1331275.

[24] F. Liu, D. Janssens, G. Wets, and M. Cools, "Annotating mobile phone location data with activity purposes using machine learning algorithms," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 3299–3311, Jun. 2013, doi: 10.1016/j.eswa.2012.12.100.

[25] Y. Lu and L. Zhang, "Imputing trip purposes for long-distance travel," *Transportation*, vol. 42, no. 4, pp. 581–595, Jul. 2015, doi: 10.1007/s11116-015-9595-0.

[26] M. Janzen, M. Vanhoof, K. W. Axhausen, and Z. Smoreda, "Estimating long-distance travel demand with mobile phone billing data," in *Proc. 16th Swiss Transp. Res. Conf.*, Monte Verità, Ascona, Switzerland, 2016, pp. 17–32.

[27] G. Han and K. Sohn, "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model," *Transp. Res. B, Methodol.*, vol. 83, pp. 121–135, Jan. 2016, doi: 10.1016/j.trb.2015.11.015.

[28] F. Wu and Z. Li, "Where did you go: Personalized annotation of mobility records," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 589–598.

[29] Z. Zhu, U. Blanke, and G. Tröster, "Inferring travel purpose from crowd-augmented human mobility data," in *Proc. 1st Int. Conf. IoT Urban Space*, 2014, pp. 44–49.

[30] J. Bao, C. Xu, P. Liu, and W. Wang, "Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests," *Netw. Spatial Econ.*, vol. 17, no. 4, pp. 1231–1253, Dec. 2017, doi: 10.1007/s11067-017-9366-x.

[31] P. Wang, Y. Fu, G. Liu, W. Hu, and C. Aggarwal, "Human mobility synchronization and trip purpose detection with mixture of hawkes processes," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 495–503.

[32] P. Wang, G. Liu, Y. Fu, Y. Zhou, and J. Li, "Spotting trip purposes from taxi trajectories," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, pp. 1–26, Feb. 2018, doi: 10.1145/3078849.

JIANPEI QIAN received the B.S. degree in traffic engineering from Southeast University, Nanjing, China, in 2013. He is currently pursuing the Ph.D. degree in transportation planning and management with Beijing Jiaotong University, Beijing, China. His research interests include travel demand modeling, travel behavior analysis, and the application of NLP in transportation.

CHUNFU SHAO received the B.S. degree in automotive engineering from Chang'an University, Xi'an, China, in 1982, and the M.S. and Ph.D. degrees in mathematical engineering from Kyoto University, Kyoto, Japan, in 1988 and 1991, respectively.

He is currently a Chief Professor of the discipline of transportation planning and management with Beijing Jiaotong University. His research interests include transportation network analysis and modeling, traffic flow modeling and prediction, and traffic assignment method.

CHUNJIAO DONG received the B.S. degree in automotive engineering from the Liaoning University of Technology, Jinzhou, China, in 2005, the M.S. degree in traffic information engineering and control from Jilin University, Changchun, China, in 2007, and the Ph.D. degree in transportation planning and management from Beijing Jiaotong University, Beijing, China, in 2011.

She is currently a Professor with Beijing Jiaotong University. Her research interests include traffic flow theory and traffic safety using statitics and machine learning.

SHICHEN HUANG is currently pursuing the Ph.D. degree in transportation planning and management from Beijing Jiaotong University, Beijing, China. His Ph.D. research includes deep learning, especially GAN and NLP, in the recognition and reconstruction of road traffic networks.