CIRRELT

CIRRELT-2019-47

Inferring Trip Destinations in Transit Smart Card Data Using a Probabilistic Topic Model

Zhanhong Cheng
Martin Trépanier
Lijun Sun

October 2019

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél. : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél. : 1-418-656-2073
Télécopie : 1-418-656-2624

Inferring Trip Destinations in Transit Smart Card Data Using a Probabilistic Topic Model

Zhanhong Cheng[1,2], Martin Trépanier[1,3], Lijun Sun[1,2,*]

1. Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)
2. Department of Civil Engineering and Applied Mechanics, Macdonald Engineering Building, 817 Sherbrooke Street West, Room 492, Montréal, Canada H3A 0C3
3. Department of Mathematics and Industrial Enginnering, Polytechnique Montréal, 2500, chemin de Polytechnique, Montréal, Canada H3T 1J4

Abstract. Inferring trip destination in smart card data with only tap-in control is an important application. Most existing methods estimate trip destination based on the continuity of trip chains, while the destinations of isolated/unlinked trips cannot be properly handled. We address this problem with a probabilistic topic model. A three-dimensional Latent Dirichlet Allocation (LDA) model is developed to extract latent topics of departure time, origin, and destination among the population; each passenger's travel behavior is characterized by a latent topic distribution defined on a three-dimensional simplex. Given the origin station and departure time, the most likely destination can be obtained by statistical inference. Furthermore, we propose to represent stations by their rank of visiting frequency, which transforms divergent spatial patterns into similar behavioral regularities. The proposed destination estimation framework is tested on Guangzhou Metro smart card data, in which the ground-truth is available. Compared with benchmark models, the topic model not only shows increased accuracy but also captures essential latent patterns in passengers' travel behavior. The proposed topic model can be used to infer the destination of unlinked trips, analyze travel pattern, and passenger clustering.

Keywords. Public transit, smart card data, destination inference, topic model, passenger clustering.

Acknowledgements. This research is funded by NSERC, Mitacs and exo (https://exo.quebec/en).

---

* Corresponding author: lijun.sun@mcgill.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2019

1 Introduction

Origin and Destination (OD) Matrix is an essential input for transit planning and operation. Most transit agencies have been relying on travel surveys to collect representative OD information. However, conducting such a survey with reasonable scale is not only costly but also time-consuming. With the recent advances of intelligent transportation systems, researchers and practitioners have started taking advantage of the transit operation data and smart card data for better planning and operation practices (1).

Smart card systems are initially designed for the purpose of Automatic Fare Collection (AFC). When the system has both tap-in and tap-out controls (e.g., using a distance-based transit fare scheme), the full itinerary (boarding time/station and alighting time/station) of each trip can be registered. However, most smart card systems across the world adopt a single fare scheme with only tap-in validation, and the alighting information (time/station) is essentially unknown. Inferring the alighting stations is a crucial problem in obtaining the OD matrix from these smart card systems.

Trip destination estimation in smart card data has always been a hot issue. Barry et al. (2) proposed two assumptions to address this issue: (1) the alighting station of a trip is very likely to be the boarding station of the immediate next trip; (2) the last alighting station of a day is usually the first boarding station of the same day. This type of "rule-based" model soon became the workhorse algorithm for smart card destination estimation. Depending on the data, current algorithms can obtain around 60% to 85% trips' destinations; these trips are often called linked trips in the literature, and the rest un-inferred trips are referred to as unlinked trips. Without the information from consecutive trips, the destination estimation of unlinked trips is more challenging. Existing methods address this problem by seeking similar trips in the passenger's historical trips; we refer them as individual-history-based model. Such as He and Trépanier (3) used the spatial and temporal kernel density probability of passengers' trips and get an additional 10% estimation for unlinked trips.

One of the drawbacks for individual-history-based model is that it requires a specific model for each individual, and all the information for training the model is from this passenger. This paper attempts to build an integrated model for all passengers and the estimation of a destination utilizes information not only from this passenger but also similar travelers. We establish a probabilistic topic model for smart card data by making an analogy with the Latent Dirichlet Allocation (LDA) model (4). Similar to the two-dimensional LDA developed by Sun et al. (5), where the topic model was used to detect anomaly in travel behavior, we extend this model to a three-dimensional LDA to capture the smart card trips. Every passenger is characterized by a latent topic distribution and the whole population share the identical topic-word distributions for departure time, origin and destination. To share more information among different passengers, we represent each station by its order in each passenger's visiting frequency, as against to directly using station ID. A case study is performed on Guangzhou Metro data, where the tap-out data is used as the ground truth to test different models. Results show our topic model has slightly higher accuracy compared with individual-history-based model, and the latent topic distributions excellently summarize the spatio-temporal travel pattern of different types of passengers. Finally, we show an application of the topic model in passenger clustering.

The remainder of the paper is organized as follows. Section 2 briefly reviews the current research on smart card data destination inference and the application of topic model in travel behavior mining. Section 3 elaborates the topic model for destination inference. The model evaluation, interpretation and the passenger clustering will be shown by a case study in section 4. Conclusions and discussions are summarized in section 5.

2 Literature Review

2.1 Destination Inference in Smart Card Data

Destination inference is an important problem in smart card data. Existing methods primary take advantage of the continuity of trip chains, and infer the destinations based on assumptions or rules. In a very first study, Barry et al. (2) proposed that the destination of a trip can be inferred by the origin of the immediate next trip, and they assumed the last destination of a day is often the first origin in the same day. Since then, many refined models have been proposed based on similar assumptions. Trépanier et al. (6) imposed a distance constraint between consecutive trips, and they further assumed the last destination of a day can also be inferred by the first origin in the next day. Munizaga and Palma (7) proposed to use generalized time instead of distance in destination inference. Further, Sánchez-Martínez (8) constructed a generalized disutility minimization objective to determine the paths and transfers between the origin and destination. Research based on similar rule-based methodology has become the mainstream, and more research can be found in (9,10,11,12). Depending on the data, the rule-based method can accomplish around 60% to 85% of the destinations; trips of which the destinations can be inferred by the rule-based model are often called linked trips.

For the O-D of unlinked trips, whose destination cannot be inferred by rule-based model, one treatment is to scale the O-D of linked trips by some methods, such as (7,13). This approach assumes the destination distribution of unlinked trip at each origin is that same with the linked trips, which is unverified. On the other hand, the destinations of unlinked trips can be estimated by historical similar trips (individual-history-based model), similar to supervised learning with labeled data. Such as Trépanier et al. (6) define a similar trip as a trip on the same route with similar departure time in the previous several days. He and Trépanier (3) used spatial and temporal kernel density probability estimated by historical trips to infer the destination of unlinked trips. Zhang et al. (14) conducted an interesting study, where a collaborative space alignment framework was presented to reconstruct smart card trips. Jung and Sohn (15) attempted to use deep learning to infer trip destinations. The result is promising, while the large amount of labeled destinations are essentially unavailable for real tap-in-only system.

In summary, existing research has developed various algorithms based on the trip continuity feature to estimate the destination of linked trips. The destination estimation of unlinked trip relies on historical similar trips. This paper provides a whole new approach to infer the destination of unlinked trips by a topic model. The proposed model is not only a prediction model, but also a generative model that captures individuals' behavioral patterns.

2.2 Topic Model in Travel Pattern Mining

Topic model is a type of statistical model initially used for extracting latent topics from a collection of documents. In recent years, the field of travel behavior research has seen an increasing trend of

applying topic model to discover meaningful latent representations from human's mobility data. For example, Goulet-Langlois et al. (16) used principal component analysis (PCA) to extract eigen-patterns from passengers' multi-week activity sequences. Sun and Axhausen (17) applied a probabilistic tensor factorization to smart card transactions to understanding urban mobility pattern. Zhao et al. (18) used smart card data to predict the time, origin and destination of next trip by a *n*-gram model.

The topic model applied in this paper is an extension of Latent Dirichlet Allocation (LDA) (4). LDA has been widely used for behavior pattern mining. Hasan and Ukkusuri (19) classified individuals' activity pattern by applying LDA to geo-location data collected from Twitter. Fan et al. (20) applied LDA to mobile phone call data, and further developed a Hidden Markov Model for complete missing mobility data. One drawback of these two models is that the spatial and temporal features are combined to one dimension, which ont only increases the vocabulary size but also loses the spatiotemporal correlation. Sun et al. (5) developed a two-dimensional LDA on license plate recognition data, where the spatial and temporal topics are modeled separately, and their interactions are characterized in a two-dimensional simplex. We applied the same methodology as (5) and extend it to smart card data with three-dimensional features (origin, destination, and time).

3 Methodology: Topic model for destination inference

This section details the probabilistic topic model for trip destination inference in smart card data. The objective is to infer the unknown trip destination in a tap-in-only system. A large portion of the destinations of linked trips could be inferred by rule-based models (2,6,7). We can train the proposed topic model by those trips with complete itineraries (although with not perfectly accurate). Next, the destinations of unlinked trips could be inferred by the trained topic model.

3.1 Model Formulation

A smart card trip could be characterized by a three-element tuple (w^t, w^o, w^d) representing the departure time, origin, and destination; where w^t is assumed to be a discrete variable in one-hour intervals. Then, all the historical trips of a passenger u can be represented as $\mathbf{w}_u = \{(w_i^t, w_i^o, w_i^d): i = 1, \dots, N_u; w_i^t \in \{1, \cdots, T\}; w_i^o, w_i^d \in \{1, \cdots, S\}\}$; where N_u is the total number of trips for passenger u, T is the number of possible departure hours, and S is the number of boarding/alighting locations. By making the analogy to the topic model in NLP, we treat each trip (w^t, w^o, w^d) as a word and $\mathbf{w}_u$ as a document (a bag of words); thus, each passenger's trips are characterized by a mixture of latent topics.

The major difference between our model and LDA is that the words here are tuples with three-dimensional attributes, which brings a problem of how to properly define latent topics over multiple dimensions. A common solution is to combine different attributes into one dimension with the vocabulary size of $T \times S \times S$, such as in (19,20). The main drawback of this approach is that it considerably increases the vocabulary size, while the new combined words are sparse with many unobserved/unlikely trips. Moreover, the interdependency between original attributes is lost (e.g., two trips with the same origin and destination but different time can become unrelated words). To address this problem, we apply a similar probabilistic tensor factorization approach as (5,17). By increasing the dimension of latent topics, we have three types of topic-word distributions, which avoids the large vocabulary set and captures interdependencies of different types of words in the latent space.

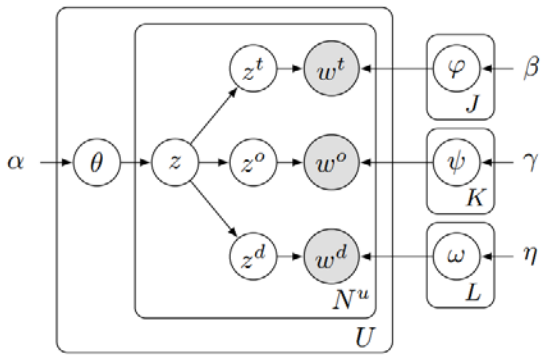

***Figure 1**: Plate notation for the graphical model.*

The latent topic is organized as a three-dimensional tensor $\mathcal{Z} \in \mathbb{R}^{J\times K\times L}$, where J, K, and L are the number of latent topics of time, origin, and destination respectively. The element $z_{j,k,l}$ of tensor $\mathcal{Z}$ corresponds to the j^{th} temporal topic z_j^t, the k^{th} origin topic z_k^o, and the l^{th} destination topic z_l^d. Each passenger's trips are characterized by a Multinomial distribution over latent topics $\mathcal{Z}$, parameterized by θ_u. Given a latent topic $z_{j,k,l}$, the topic-word distributions for departure time, origin, and destination are Multinomial distributions parameterized by φ_{z^t}, ψ_{z^o}, and ω_{z^d} respectively. The overall picture of the model can be clearly depicted by a graphical model shown in Figure 1; where α, β, γ, and η are parameters for Dirichlet priors; U is the number of passengers. We describe the generative process in Figure 1 follows:

- Draw topic distribution for each passenger $\theta_u \sim \text{Dirichlet}_{J\times K\times L}(\boldsymbol{\alpha})$.
- Draw topic-time distribution for each time topic $\varphi_j \sim \text{Dirichlet}_J(\boldsymbol{\beta})$.
- Draw origin distribution for each origin topic $\psi_k \sim \text{Dirichlet}_K(\boldsymbol{\gamma})$.
- Draw destination distribution for each destination topic $\omega_l \sim \text{Dirichlet}_L(\boldsymbol{\eta})$.
- For each passenger u, for each trip record:
  - Draw latent topic $z \sim \text{Multinomial}(\theta_u)$.
  - Obtain z^o, z^d, and z^t by z.
  - Draw $w^t \sim \text{Multinomial}(\varphi_{z^t})$.
  - Draw $w^o \sim \text{Multinomial}(\psi_{z^o})$.
  - Draw $w^d \sim \text{Multinomial}(\omega_{z^d})$.

3.2 Model Inference

The model inference involves estimating the parameters for latent topic distribution of each passenger and the topic-word distribution of each topic. In the generative process, each trip is generated from a latent topic z, which is unobserved. We use a collapsed Gibbs sampling algorithm (21) to iteratively sample the topic for each trip by the conditional probability shown in Equation (1):

$$P\left(z_i^t = j, z_i^o = k, z_i^d = l \middle| w_i^t = t, w_i^o = o, w_i^d = d, \mathbf{z}_{-i}^t, \mathbf{z}_{-i}^o, \mathbf{z}_{-i}^d, \mathbf{w}_{-i}^t, \mathbf{w}_{-i}^o, \mathbf{w}_{-i}^d\right) \propto \frac{N_{z^t=j}^{w^t=t} + \beta}{N_{z^t=j} + T\beta} \times \frac{N_{z^o=k}^{w^o=o} + \gamma}{N_{z^t=k} + S\gamma} \times \frac{N_{z^d=l}^{w^d=d} + \eta}{N_{z^t=l} + S\eta} \times \frac{N_{z^t=j,z^o=k,z^d=l}^{u} + \alpha}{N^u + JKL\alpha}. \quad (1)$$

Where $\mathbf{w}_{-i}^{(\cdot)}$ and $\mathbf{z}_{-i}^{(\cdot)}$ are trip attributes and latent topics for all other trips except trip i; $N_{(\cdot)}^{(\cdot)}$ denotes the number of trips that satisfy the condition listed in the subscript and the superscript, note that the current trip i is excluded when counting N.

The sampling procedure will converge after sufficient iterations, by then we can estimate the parameters in topic distributions and topic-word distributions by Equation (2):

$$\begin{aligned} \varphi_{t,j} &= \frac{N_{z^t=j}^{w^t=t} + \beta}{N_{z^t=j} + T\beta}, \\ \psi_{o,k} &= \frac{N_{z^o=k}^{w^o=o} + \gamma}{N_{z^t=k} + S\gamma}, \\ \omega_{d,l} &= \frac{N_{z^d=l}^{w^d=d} + \eta}{N_{z^t=l} + S\eta}, \\ \theta_{u,j,k,l} &= \frac{N_{z^t=j,z^o=k,z^d=l}^{u} + \alpha}{N^u + JKL\alpha}. \end{aligned} \quad (2)$$

3.3 Destination Inference and Ranked Locations

Having estimated all the parameters in the model, we can infer the missing destinations for trips with only origin and departure time observed. According to the Bayes' theorem, the probability for passenger u alighting at a location d given the departure time t and boarding location o takes the form:

$$\begin{aligned} P(w^d = d | w^t = t, w^o = o; u) &\propto P(w^t = t, w^o = o, w^d = d; u) \\ &= \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} P\left(w^t = t \middle| z_j^t\right) P(w^o = o | z_k^o) P\left(w^d = d \middle| z_l^d\right) P\left(z_j^t, z_k^o, z_l^d; u\right). \end{aligned} \quad (3)$$

Next, the most likely destination of a trip is the one takes the highest probability in Equation (3).

By now we have shown the complete topic model for destination inference, but there is a final impediment that prevents the model from giving a good destination estimation—the giant heterogeneity among passengers' spatial patterns. Studies have shown the frequency of individuals' historical locations follows Zipf's law (22), indicating most of the trips of a passenger are between several frequently visited locations. However, the "frequent locations" vary greatly among passengers, which means a very large number of spatial topics is required to capture the spatial heterogeneity of the entire population. The large latent space not only fails to capture the common feature among individuals, but also increases the number of unknown parameters.

To decrease the size of the latent space, we do not use unique IDs for stations, instead, we label locations by the frequency-rank in each passengers' historical trips. Specifically, denote r_u^i to be the

rank (by the order of visiting frequency) of station s_i in all the historical origins of passenger u (we only use origin in the ranking as the real destinations are unknown and the frequency of destination is roughly the same with the origin if a passenger uses smart card to and from). We transform each passenger's visited locations into the rank representation and store a mapping function $M_u(r_u^i) \rightarrow s_i$ in order to restore real stations. By doing this, and the divergent spatial patterns are essentially transformed into similar behavioral regularities (e.g., travel from the most visited station to the second most visited station). The same-ranked location for different passengers' does not correspond to the same real stations, but represents a similar degree of importance of these stations to these passengers. We build the topic model and infer the destination in the ranked reference; the estimation for real destination is then retrieved by the mapping function M_u.

4 Case Study

In this section, we examine our topic model by Guangzhou Metro smart card data. Guangzhou Metro is a tap-in and tap-out system with both origin and destination registered. We don't use any destination information in the inference process, and then validate our estimation by the real destination. We first use a rule-based model (6) to obtain all destinations of linked trips. Then, the topic model is trained by those linked trips to infer the destinations of unlinked trips. Meanwhile, the performance of the topic model is compared with four benchmark models using individuals' similar historical trips. Note that we do not consider the transfer station (not registered in the data) and only care about the two end locations of a trip.

There are in total 159 stations and the time scope of the data is three months from July 1st to September 30th, 2017. The operation time of the Metro system is 19 hours from 5:00 to 24:00, and we treat each hour as a departure time without distinguishing weekday and weekend. Among those taking more than 20 trips in the three months, we randomly select 3000 passengers to test our model. The total number of selected trips is 200,670, which means on average each person took 67 trips in the three months. Although we only train the model on selected passengers, the trained model can be easily applied to those passengers not in the training set by Gibbs sampling (5).

4.1 Model Settings

The rule-based model that we applied to infer the destinations of linked trips is similar to (6):

- Rule 1: predict the destination as the origin of the next trip in the same day.
- Rule 2: predict the last destination of a day as the first origin of the same day.
- Rule 3: predict the last destination of a day as the first origin of the next day.

The next rule will be only applied when the previous rule is not applicable to a trip. Note that any two Metro stations can be connected by transfer; therefore, we do not need to verify whether the origin of the next trip is in the vicinity of the first Metro line, which is different to the bus network in (6).

For the benchmark models, the destination is predicted as the most visited destination in this passenger's historical similar trips (in those inferred linked trips). The four benchmark models are distinguished by their different definitions for "similar trips". The four kinds of "similar trips" are defined as follows:

- (SO) Trip with the same origin.
- (ST) Trip with the same departure time (one-hour interval).
- (SOT_O) Trip with the same origin and departure time, if no such trip, use SO.
- (SOT_T) Trip with the same origin and departure time, if no such trip, use ST.

If no similar trips are found, the destination is predicted by the most visited destination by this passenger.

For our topic model, as discussed in section 3.3, instead of station IDs, we train our model by each passenger's rank of stations. Figure 2 (a) shows a passenger's number of visits per station (including both boarding and alighting). It can be found that the vast majority of trips occurs in the first several most-visited stations. Casually visited destinations, like those after rank 5, are not this passenger's major activity and are essentially hard to be predicted by any probability-based models. Figure 2 (b) shows the histogram for the number of different stations visited by each passenger. We can find most passengers visited between 5 to 20 different Metro stations in the three-month period; the number of people who visited more than 20 stations tails off. Therefore, we cut off the frequency-rank at 20, marking all stations ranked larger than 20 as 20. By doing this, each passenger's spatial vocabulary size is aligned at 20. Because the possibility of choosing cut stations is very low, as long as the cut-off point is not too small, the choice of cut-off point has little effect to the performance of our model. Representing stations by rank significantly decreases the number of latent topics needed on the spatial dimension. We perform grid search over $J = [3,4,5]$ and $K, L = [2,3,4,5]$ and select the best configuration by prediction accuracy. We finally choose $J = 4$ and $K = L = 3$, as more topics does not significantly improve the accuracy.

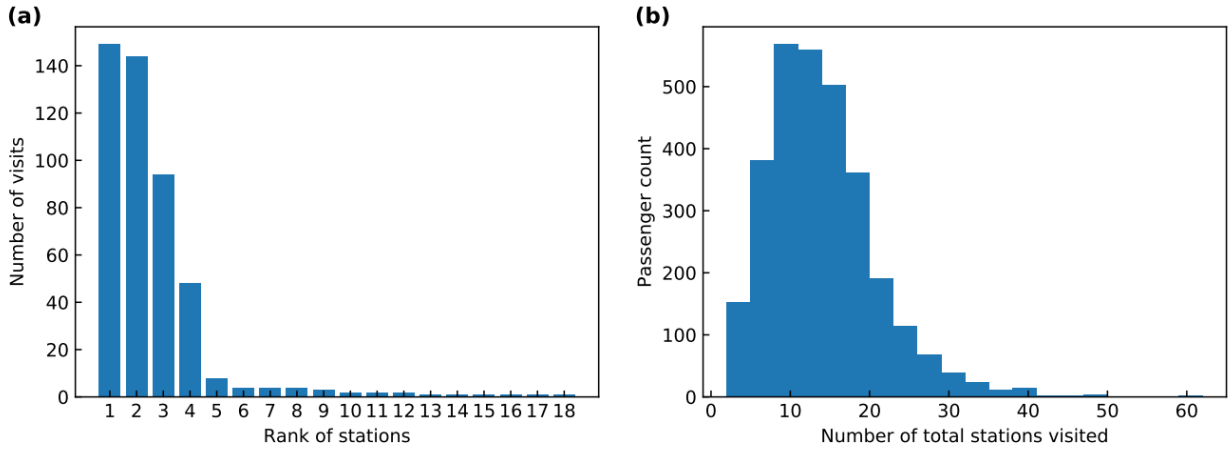

Figure 2: *Analysis of visiting frequency (a) The number of visits per stations of a passenger (including both boarding and alighting), in the order of descending frequency. (b) The histogram for the number of different stations visited by each passenger, in the 3000 passengers.*

4.2 Model Performance

The topic model cannot be trained with only origin and departure time. Therefore, we first use rule-based model to infer all the destinations of linked trips as a training set. The accuracy and the coverage of the three rules are shown in Table 1. Although the assumptions of these rules have been indirectly verified by cordon count data (2) and survey data (2,23), few study examines these

workhorse assumptions by ground-truth destinations. We can tell from Table 1 that Rule 1 using the consecutive trips could reach 87% accuracy. Although destinations inferred by Rule 2 and Rule 3 are less reliable, they are indispensable parts for the training set, because they represent the other side of passengers' travel pattern (e.g., returning home at night). The three rules together handle 85.27% trips.

***Table 1**: The destination inference accuracy and coverage of each model.*

	Coverage	Cumulative coverage	Method	Accuracy
Linked trips	44.43%	44.43%	Rule 1	87.04%
	35.48%	79.91%	Rule 2	77.65%
	5.36%	85.27%	Rule 3	60.32%
Unlinked trips	14.73%	100.00%	SO	48.68%
			ST	42.22%
			SOT_O	48.32%
			SOT_T	47.59%
			Rank topic	50.53%[a]
			No-rank topic	31.14%[b]

[a] Topic numbers $J = 4$ and $K = L = 3$. Shown by mean, std $= 0.14\%$ in 50 runs.
[b] Topic numbers $J = 5$ and $K = 10, L = 100$.

We then infer the destinations of unlinked trips by our rank-based topic model, together with four benchmark models and a topic model without rank processing, the results are shown in Table 1. The destination estimation of unlinked trip is not as accurate as linked trip (partially because of lacking ground truth). The best benchmark model is SO, with 48.68% accuracy. Our rank-based topic model shows a slight advantage over the benchmark models and is the only model reaches over 50% accuracy. Further, using the rank of stations in the topic model improves the accuracy by 20% compared to using station ID directly.

4.3 Interpret Latent Topic

By looking at the distribution over time, origin, and destination under each latent topic, we can endow semantic meanings to latent topics. The topic-word distributions are shown in Figure 3; each topic represents a type of travel behavior characterized by its word distribution. From Figure 3 (a), we can find topic T3 and T2 have very high probabilities of traveling in the morning, could be interpreted as early and late morning peaks topics respectively. Contrarily, topic T1 indicates trips in the night and T4 takes the rest of the day. For spatial topics shown in Figure 3 (b) and (c), it can be found that O1 and D2 take near 1 probability for the ranked 1st station, representing boarding and alighting at the most visited station respectively. Meanwhile, O2 and D1 represents boarding and alighting at the second most visited station. For the third spatial topics O3 and D3, the probabilities peak at the ranked 3rd station and then gradually tail off.

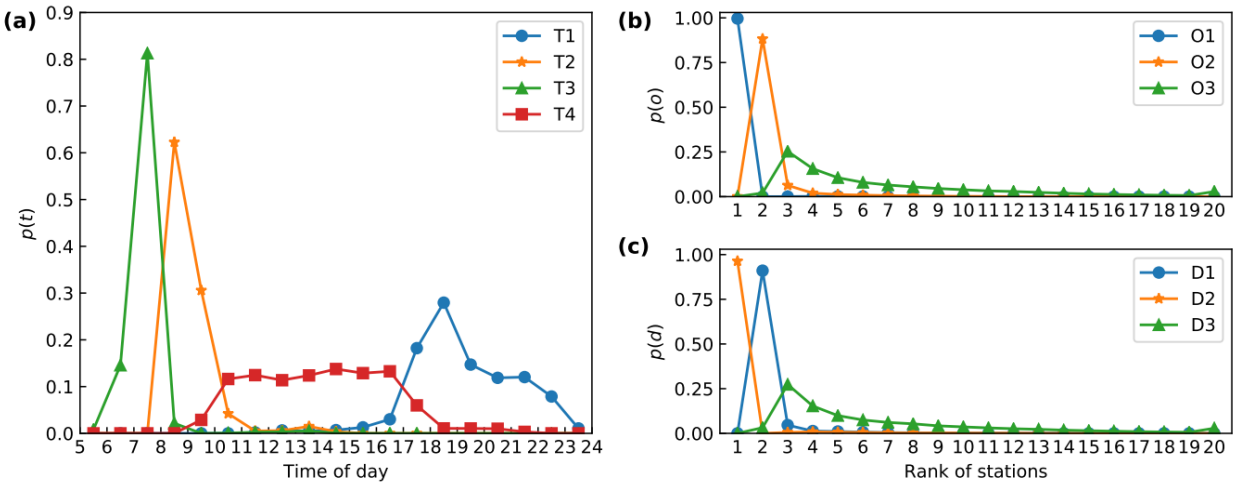

***Figure 3**: Topic-word distributions. (a) The departure time distributions of four time topics. (b) The origin distributions of three origin topics. (c) The destination distributions of four destination topics.*

It is worth mentioning that although the probability of alighting at a station after rank 3 is not zero, it is impossible to predict the destination of a trip as a station ranked after 3 by Equation (3). Because Equation (3) always predicts the destination as the most likely one, which is always the most likely destination (peak) in a particular latent destination topic. This limitation of our topic model also results in the accuracy of ranked 3rd destination being compromised by stations after rank 3. Luckily, the first two destinations make up the majority.

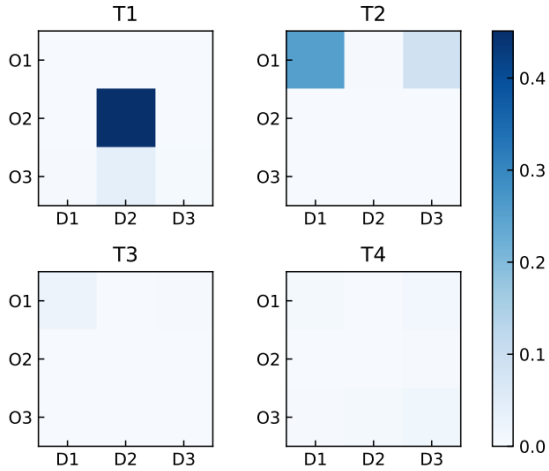

***Figure 4**: The latent topic distribution of a passenger.*

Next, we show how the latent pattern can be used to interpret passengers' travel pattern. Figure 4 shows the latent topic distribution of a passenger; each matrix represents the probabilities over origin and destination topics under a time topic. Although we don't know the exact mapping relation between the rank of a station and its real function (e.g., home/work), we can easily understand these travel patterns by common sense. It is conspicuous that there are two latent topics with significantly higher probability, indicating a possible commuting pattern. The most significant latent topic is [T1, O2, O2]; according to the semantic meaning shown in Figure 4, [T1, O2, D2] represents this passenger frequently boards at the second most visited station and alights at the most visited station in the night, indicating a possible work-home behavior. Similarly, the second significant

topic [T2, O1, D1] represents traveling from the most visited station to the second most visited location in the morning, which could be the home-work trip. Other topics have relatively low probabilities, but also reflect certain activity patterns, such as [T2, O1, D3] could be a home-activity trip in the morning and [T1, O3, D2] could be a home-activity trip in the night. Further, we can find this passenger often use Metro in the late morning (T2) and night (T1), but seldom travel in the noon and afternoon (T4).

4.4 Passenger Clustering

The latent topic distribution characterizes passengers' travel pattern, which is an excellent feature for passenger clustering. Jensen-Shannon divergence (JSD) is a metric of measuring the similarity between two probability distributions. Similar to (5), we apply the square root of the JSD as the distance for clustering.

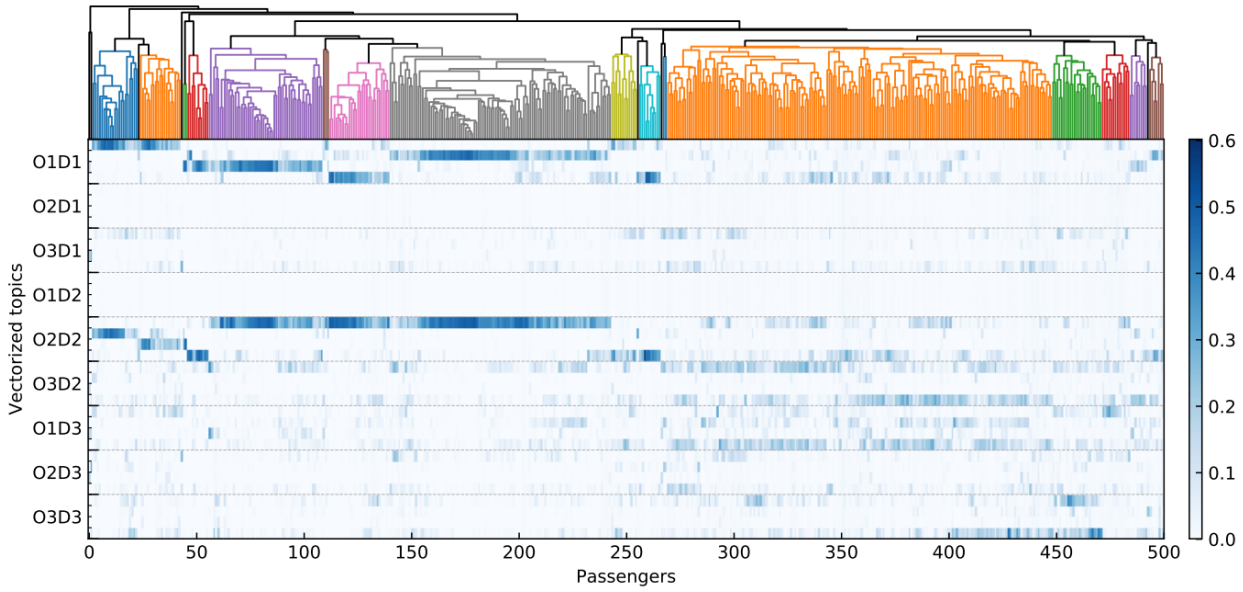

***Figure 5**: The hierarchical clustering of 500 passengers by latent topic distribution.*

The hierarchical clustering of 500 passengers by their latent topic distribution is in shown in Figure 5. The vertical axis is vectorized latent topics, where the main ticks and tick labels represent different combinations of origin and destination topics, and the minor ticks distinguish different time topics. Interestingly, passengers on the left half of the figure (around 0 to 265) show distinct two travel patterns, one form O1 to D1 and the other from O2 to D2. More specifically, the time topic of O1D1 and O2D2 are different in each cluster, showing these passengers regularly leave from a place at a certain time and then come back at another time, the time at which passengers leave and back distinguishes different clusters. On the other hand, passengers on the right half of Figure 5 (around 265 to 500) do not have an as significant commuting pattern as those on the left part, and therefore corresponds to non-commuters. The latent topic distributions of non-commuters do not concentrate at the first two location topics and show more diverse interactions between different topics. Note that we only rank the stations by the boarding stations in order to imitate the tap-in only system. If the objective is to understand passengers' travel patter, we can use both the boarding and alighting locations in the ranking and introduce more location topics to enable more refined classifications for non-commuters.

There has been a large body of research uses smart card data for passenger clustering and travel pattern mining, the feature used in our clustering is unique in capturing the compact spatial and

temporal patterns. Most existing methods capture either spatial or temporal features, such as (24,25,26,27). Ma et al. (28) clustered passengers based on spatial and temporal features; the spatial pattern is defined by a passenger repeatedly visits the similar places on a multi-day basis, the temporal pattern is defined by a passenger repeatedly starts or ends a trip in the same (close) time of a day. But the two kinds of features are independently defined and then combined together. Our latent topic distribution, instead, further captures the spatio-temporal interactions.

5 Conclusion and Discussion

This paper addresses the destination estimation problem in smart card data by a probabilistic topic model. We establish a three-dimensional LDA model than captures the time, origin, and destination attributes in smart card trips. Moreover, we introduce a station-to-rank technique that diminishes spatial divergence among passengers to discover more meaningful latent spatial topics. Our model could be used to infer the destination of unlinked trips by using the linked trips as a training set. The case study of Guangzhou Metro shows our model is comparable to individual-history-based model with a slight (2%) improvement. More than a prediction model, the proposed topic model is also a generative model that explains the probability of a trip by the individual's latent topics. Therefore, the proposed topic model can be used for travel pattern analysis and passenger clustering.

There are many future research questions. Firstly, a better representation for location/spatial data is worth investigating. We introduce the rank representation for stations; this to some extent changes the spatial information to behavior information (e.g., the first ranked station is very likely to be home) and loses interesting spatial characters. Secondly, passengers' behavior changes over time. Therefore, how to transform the topic model to a time-varying version is an interesting direction, such as (20). Finally, similar to (29), we can include extraneous variables to improve prediction accuracy.

Acknowledgement

This research is funded by NSERC, Mitacs and exo (https://exo.quebec/en).

Author Contributions

The authors confirm contribution to the paper as follows: L.S., Z.C. and M.T designed the research; L.S. and Z.C. performed the research; L.S. and Z.C. analyzed the data; L.S., Z.C., and M.T. wrote the paper. All authors reviewed the results and approved the final version of the manuscript.

References

[1] Pelletier, M.-P., M. Trépanier, and C. Morency, Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies, Vol. 19, No. 4, 2011, pp. 557–568.

[2] Barry, J. J., R. Newhouser, A. Rahbee, and S. Sayeda, Origin and destination estimation in

New York City with automated fare system data. Transportation Research Record:, Vol. 1817, No. 1, 2002, pp. 183–187.

[3] He, L. and M. Trépanier, Estimating the destination of unlinked trips in transit smart card fare data. Transportation Research Record, , No. 2535, 2015, pp. 97–104.

[4] Blei, D. M., A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation. Journal of Machine Learning Research, Vol. 3, No. Jan, 2003, pp. 993–1022.

[5] Sun, L., X. Chen, Z. He, and L. F. Miranda-Moreno, Pattern discovery and anomaly detection of individual travel behavior using license plate recognition data. In Transportation Research Board 98th Annual Meeting, 2019.

[6] Trépanier, M., N. Tranchant, and R. Chapleau, Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems, Vol. 11, No. 1, 2007, pp. 1–14.

[7] Munizaga, M. A. and C. Palma, Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies, Vol. 24, 2012, pp. 9–18.

[8] Sánchez-Martínez, G. E., Inference of public transportation trip destinations by using fare transaction and vehicle location data: Dynamic programming approach. Transportation Research Record, Vol. 2652, No. 1, 2017, pp. 1–7.

[9] Zhao, J., A. Rahbee, and N. H. Wilson, Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Computer-Aided Civil and Infrastructure Engineering, Vol. 22, No. 5, 2007, pp. 376–387.

[10] Wang, W., J. Attanucci, and N. Wilson, Bus passenger origin-destination estimation and related analyses using automated data collection systems. Journal of Public Transportation, Vol. 14, 2011, pp. 131–150.

[11] Gordon, J. B., H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci, Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transportation Research Record, Vol. 2343, No. 1, 2013, pp. 17–24.

[12] Nunes, A. A., T. G. Dias, and J. F. e Cunha, Passenger journey destination estimation from automated fare collection system data using spatial validation. IEEE Transactions on Intelligent Transportation Systems, Vol. 17, No. 1, 2016, pp. 133–142.

[13] Gordon, J. B., H. N. Koutsopoulos, and N. H. Wilson, Estimation of population origin–interchange–destination flows on multimodal transit networks. Transportation Research Part C: Emerging Technologies, Vol. 90, 2018, pp. 350–365.

[14] Zhang, F., N. J. Yuan, Y. Wang, and X. Xie, Reconstructing individual mobility from smart card transactions: a collaborative space alignment approach. Knowledge and Information Systems, Vol. 44, No. 2, 2015, pp. 299–323.

[15] Jung, J. and K. Sohn, Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. IET Intelligent Transport Systems, Vol. 11, No. 6, 2017, pp. 334–339.

[16] Goulet-Langlois, G., H. N. Koutsopoulos, and J. Zhao, Inferring patterns in the multi-week activity sequences of public transport users. Transportation Research Part C: Emerging Technologies, Vol. 64, 2016, pp. 1–16.

[17] Sun, L. and K. W. Axhausen, Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transportation Research Part B: Methodological, Vol. 91, 2016, pp. 511–524.

[18] Zhao, Z., H. N. Koutsopoulos, and J. Zhao, Individual mobility prediction using transit smart card data. Transportation Research Part C: Emerging Technologies, Vol. 89, 2018, pp. 19–34.

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

[20] Fan, Z., A. Arai, X. Song, A. Witayangkurn, H. Kanasugi, and R. Shibasaki, A collaborative filtering approach to citywide human mobility completion from sparse call records. In International Joint Conference on Artificial Intelligence, 2016, pp. 2500–2506.

[21] Griffiths, T. L. and M. Steyvers, Finding scientific topics. Proceedings of the National academy of Sciences, Vol. 101, No. suppl 1, 2004, pp. 5228–5235.

[22] Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi, Understanding individual human mobility patterns. Nature, Vol. 453, No. 7196, 2008, p. 779.

[23] Munizaga, M., F. Devillaine, C. Navarrete, and D. Silva, Validating travel behavior estimated from smartcard data. Transportation Research Part C: Emerging Technologies, Vol. 44, 2014, pp. 70–79.

[24] Morency, C., M. Trepanier, and B. Agard, Measuring transit use variability with smart-card data. Transport Policy, Vol. 14, No. 3, 2007, pp. 193–203.

[25] Ghaemi, M. S., B. Agard, M. Trépanier, and V. Partovi Nia, A visual segmentation method for temporal smart card data. Transportmetrica A: Transport Science, Vol. 13, No. 5, 2017, pp. 381–404.

[26] Mohamed, K., E. Côme, J. Baro, and L. Oukhellou, Understanding passenger patterns in public transit through smart card and socioeconomic data. In ACM SIGKDD Workshop on Urban Computing, 2014.

[27] Ma, X., C. Liu, H. Wen, Y. Wang, and Y.-J. Wu, Understanding commuting patterns using transit smart card data. Journal of Transport Geography, Vol. 58, 2017, pp. 135–145.

[28] Ma, X., Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies, Vol. 36, 2013, pp. 1–12.

[29] Yin, M., M. Sheehan, S. Feygin, J.-F. Paiement, and A. Pozdnoukhov, A generative model of urban activities from cellular data. IEEE Transactions on Intelligent Transportation Systems, Vol. 19, No. 6, 2017, pp. 1682–1696.