Individual Mobility Prediction in Mass Transit Systems using Smart Card Data: An Interpretable Activity-based Hidden Markov Approach

Baichuan Mo, Zhan Zhao*, Haris N. Koutsopoulos, and Jinhua Zhao

Abstract—Individual mobility is driven by demand for activities with diverse spatiotemporal patterns, but existing methods for mobility prediction often overlook the underlying activity patterns. Knowledge of activity patterns can improve the performance and interpretability of existing individual mobility models, leading to more informed policy design and better user experience in intelligent transportation systems. This study develops an activity-based modeling framework for individual mobility prediction in mass transit systems. Specifically, an input-output hidden Markov model (IOHMM) approach is proposed to simultaneously predict the (continuous) time and (discrete) location of an individual’s next trip using transit smart card data. The prediction task can be transformed into predicting the hidden activity duration and end location. Based on a case study of Hong Kong’s metro system, we show that the proposed model can achieve similar prediction performance as the state-of-the-art long short-term memory (LSTM) model. Unlike LSTM, the proposed IOHMM approach can also be used to analyze hidden activity patterns, which provides meaningful behavioral explanations while enhancing our ability to generate insightful behavioral explanations, which is useful for user-centric policy design and intelligent transportation applications such as personalized traveler information.

Index Terms—Individual mobility; Next trip prediction; Hidden Markov model; Smart card data; Public transit

I. INTRODUCTION

INDIVIDUAL mobility prediction describes the prediction of human movements over space and time at the individual level. It has important smart city and smart transportation applications, including personalized traveler information, targeted demand management, etc. Despite the emergence of extensive urban data, it is a challenging problem to accurately predict individual mobility. Travel behavior concerns multiple dimensions (most notably the temporal and spatial dimensions), exhibits longitudinal variability for an individual, and varies across individuals [1], making the mobility prediction problem difficult to tackle. This is especially challenging for public transit systems because they can only observe part of individual mobility as transit trips, typically through smart card records. The same issue is also relevant to the new app-based mobility systems, as their apps only collect partial mobility information when a user consumes their services. Such data directly generated by usage of various mobility systems are referred to as intrinsic mobility data [2].

Individual mobility prediction is complex and multi-dimensional. While the literature mostly focuses on the problem of next location prediction [3, 4, 5, 6], relatively less attention was given to the problem of next trip prediction, especially using intrinsic mobility data. In a prior related study, Zhao et al. [2] defined several sub-problems related to the next trip prediction problem in the context of mass rail transit systems based on smart card data. It is found that, while it is easier to predict whether an individual travels or not, it is much harder to predict when and where they go next. This is not surprising because of the large number of possible combinations of people’s spatiotemporal choices. It is generally challenging to deal with high-dimensional problems, especially when the data is relatively sparse at the individual level. Besides, the existing methods are limited in that the time of travel is often treated as a categorical variable. The arbitrary discretization of time does not represent people’s temporal choices adequately and may exacerbate the data sparsity issue. Furthermore, while spatial and temporal choices of travel are typically made simultaneously, existing methods often simplify the problem to a sequential prediction task [2]. This study aims to address these limitations with an activity-based approach.

Inspired by activity-based models commonly used in travel behavior research, the main objective of the paper is to develop a methodology to simultaneously predict the time and location of an individual’s next trip through their latent activity patterns. Instead of directly predicting travel behavior, we propose an input-output hidden Markov model (IOHMM) approach to analyze the underlying activity behavior. Individual mobility is driven by demand for activities with diverse spatiotemporal patterns, and thus uncovering the latter can help us predict the former. For example, the prediction of the activity duration is equivalent to that of the start time of the next trip. The specific problem formulation and model design are based on smart card data in mass transit systems, but they can be adapted to other mobility systems, e.g., ride-hailing. Transit smart card data from Hong Kong’s Mass Transit Railway (MTR) system are used to illustrate the applicability of the proposed methodology.

Baichuan Mo is with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139 USA.
Zhan Zhao is with the Department of Urban Planning and Design, The University of Hong Kong, Hong Kong, China.
Haris N. Koutsopoulos is with the Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115 USA.
Jinhua Zhao is with the Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, 02139 USA.
* Corresponding author: zhanzhao@hku.hk
The contributions of this study are summarized as follows:

- Existing individual mobility prediction models often lack natural behavioral interpretation, limiting their applicability for intelligent policy design. This paper introduces an activity-based modeling framework that captures the underlying generative mechanism of travel behavior and uncovers people’s travel purposes not directly observable in the data. To the best of our knowledge, this is the first study that adopts an activity-based model for individual mobility prediction using transit smart card data.

- We adopt an IOHMM approach for activity-based modeling, and extend it for simultaneous prediction of the discrete location and continuous time of travel. It has been shown that the temporal aspect of individual mobility is least predictable [2], partly because of the arbitrary discretization of time. In our extended IOHMM, the trip start time is represented as a continuous emission variable, and can be predicted jointly with a discrete location variable through inference of latent activity types.

- We associate the hidden states in the IOHMM with individual’s hidden activities, and propose a Gibbs sampling method to extract and visualize hidden activity patterns with semantic explanations. We also show how the estimated model parameters and transition matrix can be used for the model’s interpretability.

- The proposed methodology is demonstrated using transit smart card data from Hong Kong MTR. Compared to state-of-the-art deep learning models, the results show that our activity-based model can achieve competitive predictive performance, while offering significantly more interpretability into the underlying activity patterns and travel purposes. The combination of performance and interpretability makes our approach more versatile and actionable.

II. LITERATURE REVIEW

Demand prediction in public transit systems can be categorized into aggregate and individual levels. The aggregate demand prediction [7, 8, 9, 10, 11] has been extensively studied with the recent development of deep learning methods. A summary of previous works can be found in Fang et al. [12]. However, the individual-level prediction gains less attention compared to the aggregated ones.

The literature on individual mobility prediction mostly focuses on the problem of next location prediction, rather than next trip prediction. Most existing methods for next location prediction are based on mining sequential patterns of individual location histories. Simple Markov chain (MC) models have shown to be able to achieve good prediction performance [13, 4]. Asahara et al. [14] proposed a mixed Markov chain model (MMM) for next location prediction by identifying the group a particular individual belongs to and applying a specific MC model for that group. Mathew et al. [15] presented a hybrid method of clustering location histories according to their characteristics before training a hidden Markov model (HMM) for each cluster. Recently, due to the rapid advance of deep learning, variants on Recurrent Neural Network (RNN) models have been adopted for next location prediction, and showed improved prediction performance over MC models [16, 17, 18]. Some of the most competitive models today are based on the Long Short-Term Memory (LSTM) architecture [19, 20, 21]. Similar methods have also been proven successful for vehicle trajectory prediction problems [22, 23]. However, none of these methods explicitly consider the temporal behavior of individual mobility, e.g., when to start the next trip. This is important for any mobility service because travel demand is dynamic and time-sensitive. In addition, despite the superior predictive performance, the deep learning models are generally black-box and difficult to interpret, making them unfit as supportive tools for policy design.

For the next trip prediction, we have to model both the spatial and temporal choices of individual trip-making. Only a few prior studies have dealt with this issue. Gidófalvi and Dong [24] developed a continuous-time Markov model to predict when an individual will leave their current location and where will they go next. Hsieh et al. [25] introduced a time-aware language model, T-gram, to predict when an individual leaves a location by extracting location-specific time distributions from social media check-in data. Focusing on mass transit systems, Zhao et al. [26] explicitly formulated the spatiotemporal choices of individual passengers as a sequence of decisions, and proposed a mobility N-gram model to predict the choices associated with the next trip—the trip start time, origin, and destination. It is found that the start time is the least predictable aspect of the next trip. While the low predictability for trip start time is to some extent rooted in people’s inherent behavioral variability, the discrete representation of time in most existing models is likely to limit our ability to predict temporal behavior. The key challenge is to capture the complex interaction of continuous time choices and discrete location choices. One way to do this is through latent variables representing hidden activities between trips [26].

In the activity-based analysis of travel behavior, travel is treated as being derived from the need to pursue activities distributed in space and time [27, 28, 29]. With a more realistic representation of travel behavior, activity-based models are intuitive and interpretable, and have been widely used in transportation planning, though they usually depend on detailed manual survey data. Recent years have seen growing interests in automatically detecting activity patterns from large-scale human mobility traces [30, 31], including transit smart card data. Han and Sohn [32] developed a Continuous Hidden Markov Model (CHMM) to impute the sequence of activities for each trip chain. Zhao et al. [26] proposed a spatiotemporal topic model to discover latent activity patterns from smart card data. While these methods can enrich mobility data with behavioral semantics, none of them are suitable for predicting future trips. An activity-based approach to individual mobility prediction is needed to capture spatiotemporal correlation and enhance behavioral interpretability.

This study presents the first activity-based approach to individual mobility prediction. It is based on the Input-Output Hidden Markov Model (IOHMM), which is an extension of standard HMMS [33]. The standard HMMS assume homogeneous transition and emission probabilities, in which the
contextual information cannot be captured. To overcome this limitation, the IOHMM was proposed to incorporate additional information. Specifically, transition probabilities in IOHMMs are conditional on the input and thus depend on time. IOHMM was designed for sequence data processing, and has been applied for diverse problems including grammar inference [34], gesture recognition [35], audio processing [36], electricity price forecasting [37], and urban activity generation [38]. As we will demonstrate in Section III-C, the IOHMM architecture can be adapted to (1) capture the dynamics of individual travel-activity histories, (2) incorporate rich contextual information for improved prediction performance, and (3) jointly predict both discrete (location) and continuous (time) attributes of trips/activities. Table I summarizes the key difference of our model compared to some of the existing ones for individual mobility prediction.

**TABLE I: Summary of individual mobility prediction studies**

<table>
<thead>
<tr>
<th>Study</th>
<th>Problem</th>
<th>Data</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asahara et al. [14]; Gambs et al. [13]</td>
<td>Location Prediction</td>
<td>GPS Data</td>
<td>Markov Chain</td>
</tr>
<tr>
<td>Mathew et al. [15]</td>
<td>Location Prediction</td>
<td>GPS Data</td>
<td>HMM</td>
</tr>
<tr>
<td>Al-Molegi et al. [17]</td>
<td>Location Prediction</td>
<td>GPS Data</td>
<td>RNN</td>
</tr>
<tr>
<td>Liu et al. [16]; Kong and Wu [19]</td>
<td>Location Prediction</td>
<td>LBS Data</td>
<td>RNN / LSTM</td>
</tr>
<tr>
<td>Feng et al. [18, 21]</td>
<td>Location Prediction</td>
<td>Mobile Phone Data</td>
<td>RNN / LSTM</td>
</tr>
<tr>
<td>Gidófalvi and Dong [24]</td>
<td>Location+Time Prediction</td>
<td>GPS Data</td>
<td>Continuous-Time Markov Chain</td>
</tr>
<tr>
<td>Zhao et al. [2]</td>
<td>Location+Time Prediction</td>
<td>Transit Smart Card Data</td>
<td>Bayesian N-Gram</td>
</tr>
<tr>
<td>This Study*</td>
<td>Location+Time Prediction</td>
<td>Transit Smart Card Data</td>
<td>IOHMM</td>
</tr>
</tbody>
</table>

*Only this study adopts an activity-based approach

---

**III. METHODOLOGY**

**A. Problem description**

In this section, we illustrate the methodology based on transit smart card data for consistency with the case study. However, as mentioned in Section I, this framework applies to all intrinsic mobility data. More discussions on how to extend the model to other intrinsic mobility data are shown in Section V.

Transit smart card data is one of the most important intrinsic mobility data. It includes passengers’ tap-in and tap-out 1 transaction records, which can provide the chronological public transit (PT) trip histories of each individual. The trip structure is shown in Figure 1. Each trip starts with boarding at an origin station and ends with alighting at a destination station. The boarding (and alighting) times and locations are known from the transit smart card data. The unique ID of each smart card allows us to track the trip histories of each anonymous individual. Between two consecutive trips, a passenger may have some activities such as working, staying at home, etc. In this study, the latent behavior of an individual between two adjacent trips is referred to as a hidden activity. Different from a typical definition of an activity where passengers stay in a place, the hidden activity in this study may include unobserved trips such as taking a taxi to another place. Due to data limitations, we cannot identify people’s trips outside the PT system. Thus, we assume that no matter what people have done between two adjacent transit trips, this process is treated as a single hidden activity. The alighting station of the last trip and the boarding station of the next trip are referred to as activity start and end locations, respectively. Our goal is to predict when and where the next trip will start given a sequence of recorded trip histories.

![Fig. 1: Public transit trip structure](image)

**B. Activity-based modeling framework**

The duration of people’s hidden activities can vary greatly, anywhere from one hour (e.g. shopping) to several days (e.g. vacation). The wide range of activity duration makes it challenging to predict. In this study, we set the basic prediction interval as one day. A sequence of consecutive activities is extracted from the smart card data on a specific day and each individual may have multiple sequences of activities. The choice of prediction interval of one day not only reduces the scope of the prediction problem (from infinity to 24 hours) but also represents the basic period of regularity for human mobility and activity patterns [39, 40, 2]. Specifically, each day spans from 4:00 AM to 4:00 AM of the next calendar day, which better matches people’s daily activity schedules and the operating time of transit.

For a user \( u \), the recorded public transit trips in day \( v \) are represented as

\[
S^{u,v} = \{ (o_1^{u,v}, d_1^{u,v}, x_1^{u,v}, y_1^{u,v}), \ldots, (o_{T_u,v}^{u,v}, d_{T_u,v}^{u,v}, x_{T_u,v}^{u,v}, y_{T_u,v}^{u,v}) \}  
\]  

(1)
Therefore, our goal is to predict \( q_{t}^{u,v} \), the trajectories and other information (e.g., weather). For the first activity, we explicitly define \( d^{u,v}_{0,t} = \text{“null”} \) and \( y^{u,v}_{0,t} = 4:00 \text{ AM}. \) An example to illustrate the relationship between \( S^{u,v} \) and \( H^{u,v} \) is shown in Figure 2. After a trip ends, \( p^{u,v}_{t} \) is directly observed from the transit smart card records. Therefore, our goal is to predict \( q^{u,v}_{t} \) and \( r^{u,v}_{t} \) given historical trajectories and other information (e.g., weather).

It is worth noting that we do not consider the time period from \( y^{u,v}_{t+1} \) to 4:00 AM next day as the last activity interval because its duration is deterministic. If \( y^{u,v}_{t} \) is known. This study focuses on predicting the next trip’s time and location, but there is no corresponding next trip for this activity. Therefore, there is no need to predict the last activity, and it is excluded from further analysis.

Fig. 2: Relationship between \( S^{u,v} \) and \( H^{u,v} \). There are two trips (thus two activities) in the day for this example.

### C. IOHMM for activity prediction

IOHMM is proposed to capture exogenous contextual information over time, which allows the modeling of heterogeneous transition and emission probabilities. The structure of IOHMM for individual activity modeling is shown in Figure 3. \( A_{t} \) is the \( t \)-th hidden activity (a latent random variable) and \( z_{t} \) is a vector of observed input variables containing contextual information (e.g., weather, day of week, \( p_{t} \), etc.). The superscript \((u,v)\) is ignored for simplicity. Since each hidden activity can be encoded as a latent state in IOHMM, the IOHMM architecture matches well with the activity-based modeling framework.

The model consists of three key components: 1) initial state probability \( \pi_{i} = \mathbb{P}(A_{1} \in i \mid z_{1}; \theta_{im}) \), where \( i \in \mathcal{A} \) and \( \mathcal{A} \) is the state space; It quantifies the distribution of the first activity’s type. 2) transition probability: \( \varphi_{ij,t} = \mathbb{P}(A_{t} = j \mid A_{t-1} = i, z_{t}; \theta_{tr}) \), which quantifies the probability that next activity is \( j \) given this activity is \( i \), and 3) emission probability: \( \delta_{i,t} = \mathbb{P}(q_{t}, r_{t} \mid A_{t} = i, z_{t}; \theta_{em}) \), which quantifies the distributions of activity duration and end location. \( \theta_{im}, \theta_{tr} \), and \( \theta_{em} \) are parameters of initial, transition, and emission probability functions, respectively. The likelihood of a data sequence under this model is given by:

\[
L(\theta) = \prod_{A_{1},...,A_{T}} \mathbb{P}(A_{1} \mid z_{1}; \theta_{im}) \prod_{t=2}^{T} \mathbb{P}(A_{t} \mid A_{t-1}, z_{t}; \theta_{tr}) \\
\prod_{t=1}^{T} \mathbb{P}(q_{t}, r_{t} \mid A_{t}, z_{t}; \theta_{em})
\]

where \( \theta = [\theta_{im}, \theta_{tr}, \theta_{em}] \).

The model is estimated by the Expectation-Maximization (EM) algorithm.

**E-step:** Denote the estimated parameters at iteration \( k - 1 \) of M-step as \( \theta^{(k-1)} \) (if \( k = 1 \), use the initial values of the parameters). From \( \theta^{(k-1)} \) we can obtain the three probabilities as \( \pi^{(k-1)}_{i}, \delta^{(k-1)}_{i,t} \), and \( \varphi^{(k-1)}_{ij,t} \). Then, the forward and backward variables (denoted as \( \alpha^{(k)}_{i,t} \) and \( \beta^{(k)}_{j,t} \), respectively) are calculated as

\[
\alpha^{(k)}_{i,t} = \mathbb{P}(q_{1:t}, r_{1:t} \mid A_{1} = i \text{ and } z_{1:t}) = \delta^{(k-1)}_{i,t} \sum_{l \in \mathcal{A}} \varphi^{(k-1)}_{l,i,t} \alpha^{(k)}_{l,t-1}
\]

\[
\beta^{(k)}_{j,t} = \mathbb{P}(q_{t+1:T}, r_{t+1:T} \mid A_{t} = j, z_{1:T}) = \sum_{l \in \mathcal{A}} \varphi^{(k-1)}_{j,l,t} \beta^{(k)}_{l,t+1}
\]

where \( \alpha^{(k)}_{i,1} = \pi^{(k-1)}_{i} \delta^{(k-1)}_{i,1} \) and \( \beta_{j,T} = 1 \). The subscripts \( 1 : t \) indicates a list of the corresponding variable with subscript from 1 to \( t \). Then, we calculate the posterior state probability and posterior transition probability as:

\[
\gamma^{(k)}_{i,t} = \mathbb{P}(A_{t} = i \mid q_{1:T}, r_{1:T}, z_{1:T}) = \alpha^{(k)}_{i,t} / \sum_{i \in \mathcal{A}} \alpha^{(k)}_{i,t}
\]

\[
\xi^{(k)}_{ij,t} = \mathbb{P}(A_{t} = j, A_{t-1} = i \mid q_{1:T}, r_{1:T}, z_{1:T}) = \varphi^{(k-1)}_{j,i,t} \beta^{(k)}_{j,t} / \sum_{i \in \mathcal{A}} \alpha^{(k)}_{i,t} \beta^{(k)}_{j,t}
\]

where \( \gamma^{(k)}_{i,t} \) is the complete data likelihood at iteration \( k \), defined as \( L_{c} = \sum_{t \in \mathcal{A}} \alpha^{(k)}_{i,t} \). Obtaining \( \alpha^{(k)}_{i,t}, \beta^{(k)}_{j,t}, \gamma^{(k)}_{i,t} \), and \( \xi^{(k)}_{ij,t} \) for all \( i, j \in \mathcal{A} \) and \( t = 1, ..., T \) finishes the E-step.

**M-step:** The probability parameters in iteration \( k \) is updated
by maximizing the expected data log likelihood:

\[
Q(\theta; \theta^{(k-1)}) = \sum_{i \in A} \gamma_i^{(k)} \log \mathbb{P}(A_1 = i \mid z_1; \theta_{in}) + \sum_{t=2}^T \sum_{i,j \in A} \gamma_{ij,t}^{(k)} \log \mathbb{P}(A_t = j \mid A_{t-1} = i, z_t; \theta_{tr}) + \sum_{i=1}^T \gamma_i^{(k)} \log \mathbb{P}(q_t, r_t \mid A_{t-1} = i, z_t; \theta_{em})
\]

where \( M \) is the set of possible numbers of hidden activities. In this study, \( M = \{3, 4, \ldots, 7\} \) is used. It worth noting that we also tested other cluster quality metrics, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC). Numerical results show that the silhouette coefficient works best for determining the number of hidden activities.

**Three probability functions:** The multinomial logistic regression is used to model the initial probability and transition probability. Specifically, we have:

\[
\mathbb{P}(A_1 = i \mid z_1; \theta_{in}) = \frac{\exp(\theta_{in,i} \cdot z_1)}{\sum_{j \in A} \exp(\theta_{in,j} \cdot z_1)}
\]

where \( \theta_{in,i} \) are the coefficients of the initial state probability function at state \( i \).

\[
\mathbb{P}(A_t = j \mid A_{t-1} = i, z_t; \theta_{tr}) = \frac{\exp(\theta_{tr,ij} \cdot z_t)}{\sum_{j' \in A} \exp(\theta_{tr,j'} \cdot z_t)}
\]

where \( \theta_{tr,ij} \) are the coefficients of the state transition probability function when the next state is \( j \) given the current state is \( i \).

Note that we use the multinomial logit (MNL) model to characterize the probabilities, instead of directly estimating the individual values (like the typical HMM), because these probabilities are also determined by \( z_t \). And the linear-in-parameters structure of the MNL model facilitates the model’s interpretability. The same probability modeling methods can also be found in [38].

In terms of the emission probability, it is worth noting that \( q_t \) is a discrete random variable while \( r_t \) continuous. Given a hidden activity, we assume conditional independence between \( q_t \) and \( r_t \), that is, we assume that the correlation between \( q_t \) and \( r_t \) within the same activity type is negligible, which simplifies the model estimation.

\[
\mathbb{P}(q_t, r_t \mid A_t = i, z_t; \theta_{em}) = \mathbb{P}(q_t \mid A_t = i, z_t; \theta_{emq}) \cdot \mathbb{P}(r_t \mid A_t = i, z_t; \theta_{emr})
\]

For the activity end location distribution, a similar multinomial logistic regression model is used, where

\[
\mathbb{P}(q_t = l \mid A_t = i, z_t; \theta_{emq}) = \frac{\exp(\theta_{emq,i,l} \cdot z_t)}{\sum_{l' \in \mathcal{L}} \exp(\theta_{emq,i,l'} \cdot z_t)}
\]

where \( \theta_{emq,i,l} \) are the coefficients for emission probability of activity location where the location is \( l \) given the current state is \( i \). \( \mathcal{L} \) is the set of location candidates. For user \( u \), \( \mathcal{L}^u \) is defined as all stations that he/she has visited in the smart card data records.

In terms of the duration distribution, we assume a Gaussian distribution with the mean expressed as a linear function of explanatory variables. This formulation enables the evaluation of the impact of contextual variables on activity duration and can be estimated efficiently (like a linear regression).

\[
\mathbb{P}(r_t \mid A_t = i, z_t; \theta_{emr}) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(r_t - \theta_{emr,i} \cdot z_t)^2}{2\sigma_i^2}}
\]
where $\theta_{m, r, i}$ and $\sigma_i$ denote the coefficients and the standard deviation of the model when the hidden state is $i$. Ideally, $\sigma_i$ can also be a function of $z_t$. However, this will significantly increase the model’s estimation difficulty. Hence, we assume that under a specific activity type, the variance of duration is a constant, and different variance patterns can be captured by different activity types.

It is worth noting that the conditional independence assumption allows us to model the joint probability of $r_t$ and $q_t$ through the hidden activity. This makes it feasible to estimate the two variables simultaneously. Moreover, the conditional independence between $q_t$ and $r_t$ under the same activity type can be further justified by adding, if needed, more activity types as a simple extension of the current framework. For example, one may argue that the duration of a lunch activity depends on where they have the meal. This can be captured by separating the lunch activity into more sub-activities, where each sub-activity is associated with an identical restaurant (location). In this way, the current conditional independence framework still works.

E. Prediction formulation

The IOHMM supports predicting the next activity duration and location given today’s trajectories after training. Here we only give the formulation for predicting the duration, and that for location prediction can be derived in the same way. Observe that predicting the duration for the next activity is equivalent to obtaining $P(r_{t+1} | q_{1:t}, r_{1:t}, z_{1:t+1})$. This is because $q_{1:t}, r_{1:t}, z_{1:t+1}$ are all observed information. By the conditional independence, we have $P(r_{t+1} | q_{1:t}, r_{1:t}, z_{1:t+1}) = P(r_{t+1} | r_{1:t}, z_{1:t+1})$. By the law of total probability:

$$P(r_{t+1} | r_{1:t}, z_{1:t+1}) = \sum_{i \in A} P(r_{t+1} | A_{t+1} = i, z_t) \cdot P(A_{t+1} = i | r_{1:t}, z_{1:t+1})$$

(19)

The first term in the right hand side (RHS) of Eq. 19 is the emission probability. And the second term can be expanded as:

$$P(A_{t+1} = i | r_{1:t}, z_{1:t+1}) = \frac{P(A_{t+1} = i, r_{1:t} | z_{1:t+1})}{\sum_{j \in A} P(A_{t+1} = j, r_{1:t} | z_{1:t+1})}$$

(20)

where

$$P(A_{t+1} = i, r_{1:t} | z_{1:t+1}) = \sum_{i \in A} P(A_{t+1} = j | A_t = i, z_{t+1}) \cdot P(A_t = i, r_{1:t} | z_{1:t})$$

(21)

The first term in the RHS of Eq. 21 is the state transition probability. And the second term is essentially the forward variable (Eq. 7) when only incorporating the emission probability of duration. Therefore, based on the forward variable, state transition probability, and emission probability, one can output the distribution of $r_{t+1}$ given $r_{1:t}, z_{1:t+1}$. We adopt the value with the highest probability density as the predicted duration.

For the location distribution, we can derive $P(q_{t+1} | q_{1:t}, z_{1:t+1}, r_{1:t+1})$ using the same method. The location with the highest probability is selected as the prediction.

F. Model interpretability

The IOHMM allows us to explore the mobility patterns of an individual. In the following discussion, the subscript $t$ is ignored as we focus on deriving the general pattern over history.

Activity pattern identification: To identify the latent activity, four distributions conditioning on a specific activity label are calculated: 1) duration distribution $P(r | A = i)$, 2) end location distribution $P(q | A = i)$, 3) start time (i.e., last trip end time) $P(y | A = i)$, and 4) start location distribution $P(p | A = i)$. Since the transition matrix describes how passengers moving from one activity to another, we are also curious about $P(A_i = j | A_{i-1} = i)$ for all $i, j \in A$. Based on these distributions, we can assign a semantic label (e.g., home, work) to each hidden activity manually.

These distributions/parameters are calculated based on Gibbs sampling as illustrated in Algorithm 1. It is worth noting that we assume $P(z)$ is the same as the distribution of $z$ in historical trajectories. Hence, instead of sampling $z \sim P(z)$, we can generate samples by going through all histories (i.e., $t = 1, ..., T^{u,v}$, for $u = 1, ..., V^u$, where $V^u$ is the total number of travel days for user $u$). This can be seen as bootstrapping). The sampling process is repeated $N$ times. And the intended distributions can be directly obtained from the generated sequences (e.g., for a discrete variable, we can directly count the conditional frequency in the generated sequences).

Algorithm 1 Activity pattern identification for a user $u$ using Gibbs sampling

Input: Trained IOHMM; Trip history of user $u$

Output: Intended probability distribution of user $u$

1: Initialize the number of sampling $N$.
2: for $n = 1$ to $N$ do
3: for $v = 1$ to $V^u$ do
4: Sample $A_{1:v} \sim P(A_{1:v} | z_{1:v})$
5: Sample $r_{1:v} \sim P(r_{1:v} | A_{1:v}, z_{1:v})$ and $q_{1:v} \sim P(q_{1:v} | A_{1:v}, z_{1:v})$
6: for $t = 2$ to $T^{u,v}$ do
7: Sample $A_{t:v} \sim P(A_{t:v} | A_{t-1:v}, z_{t:v})$
8: Sample $r_{t:v} \sim P(r_{t:v} | A_{t:v}, z_{t:v})$ and $q_{t:v} \sim P(q_{t:v} | A_{t:v}, z_{t:v})$
9: end for
10: end for
11: Save the generated activity sequences and corresponding contextual information in iteration $n$ as $H^u_n$ and $z^u_n$ respectively.
12: end for
13: Obtain the intended distribution described above for user $u$ based on $\{[H^u_1, z^u_1], ..., [H^u_N, z^u_N]\}$.

Probability coefficient explanation: After training the model, we can obtain $\theta$ for each probability function. Since all functions adopt a linear relationship between $\theta$ and $z$, the value of $\theta$ enables interpretability and validation of the training results. For example, we may expect rain to have a positive effect on the duration of all activities.
IV. CASE STUDY

A. Data

The dataset used for the case study contains transit smart card records from 500 anonymous users between July 2014 and March 2017 in the Hong Kong Mass Transit Railway (MTR) system. These users are selected randomly from all individuals with at least 300 active days of transit usage during the study period, which excludes occasional users and short-term visitors such as tourists. This is because a minimum amount of personal travel history is required to achieve reasonable prediction performance. The mobility prediction for infrequent users and short-term visitors requires future research. It is worth noting that though an individual may hold more than one smart card and a smart card data may represent multiple users, we assume that each card ID corresponds to only one user [2].

We partition the personal daily activity sequences of each user into training and test sets. The test set consists of the sequences from 20% randomly selected active days. The remaining sequences form the training set. The proposed model is specified for each user based on their own training data.

B. Travel patterns

The travel patterns of selected sample individuals are shown in Figure 4. Figure 4(a) shows the distribution of the number of active days (i.e., days with at least one trip). We observe most of the samples have less than 400 active days during the 2.5 years of the analysis period. Figure 4(b) shows the distribution of the number of trips per active day. An individual typically makes two trips in a day, likely as a result of commuting to and from work. Note that only rail-based trips are considered in the case study. Thus, this distribution is an underpresentation of the true travel intensity of users. The distribution of activity duration is shown in Figure 4(c). For the first activity (i.e., the first activity in a day), we observe a prominent peak of around 2 hours, which may represent short-term leisure. For the remaining activities (i.e., all activities in a day excluding the first one), a major peak at around 10 hours is observed, which may indicate the work activity (start at 8:00 AM and end at 18:00). Another peak for remaining activity duration is seen at around 4 hours. This may represent the weekday “staying at home” activity because the start of a day is set as 4:00 AM and people usually leave home for work at around 8:00 AM. There is a sub-peak at around 14 hours for the first activity, which may correspond to the holiday “staying at home” activity where people stay at home until 18:00 and then leave home for leisure. For the remaining activities (i.e., all activities in a day excluding the first one), a major peak at around 10 hours is observed, which may indicate the work activity (start at 8:00 and end at 18:00). Another peak for remaining activity duration is around 2 hours, which may represent short-term dining/entertainment activities. The trip start time distribution is shown in Figure 4(d), as expected, a morning peak at 8:00 AM and an evening peak at 18:00 are observed.

C. Evaluation metrics and benchmark models

Recall that the proposed IOHMM can output the predicted activity duration \( r_t^{u,v} \) and end location \( q_t^{u,v} \). \( q_t^{u,v} \) is a categorical variable and the prediction accuracy is used for the performance evaluation. \( r_t^{u,v} \) is a continuous variable and the prediction accuracy is used for the performance evaluation. \( r_t^{u,v} \) is a continuous variable. The predicted \( R^2 \) (i.e. \( R^2 \) in the test data set) is used as the main performance metric because it typically ranges between 0 and 1, which is consistent with the range of prediction accuracy.

To properly evaluate the proposed IOHMM, we compare it against two types of models for benchmarking. The first group includes simple and straightforward models and can be seen as a “lower bound” of the prediction performance. The second group is based on more advanced machine learning methods that are commonly used for sequential prediction. It can be seen as providing an approximate “upper bound” of the prediction performance for existing approaches. Specifically, linear regression (LR) and the first-order Markov Chain (MC) model are used as the first type benchmark models for predicting \( r_t^{u,v} \) and \( q_t^{u,v} \), respectively. LR is used because it is the most commonly used model for continuous variable prediction. The MC model is used because it was shown in [44] that the first-order MC can approach the limit of predictability for the next location prediction and it was previously used in [2] as the baseline model for location prediction. Moreover, we also include the mobility n-gram (NG) model proposed in [2] as a benchmark for the location prediction.

The LR model for user \( u \) is formulated as

\[
r_t^{u,v} = \beta_0^{u} + \beta^u \cdot z_t^{u,v} + \epsilon^u \quad \forall v, t
\]

where \( \beta_0^{u} \) is the intercept and \( \beta^u \) is the vector of parameters to estimate. \( \epsilon^u \) is the error term.

In terms of the MC model, the distribution of the activity end location (i.e. next trip origin) is formulated as

\[
P(q_t^{u,v} | q_{t-1}^{u,v}) = \frac{C(o_t^{u,v}, q_{t-1}^{u,v}) + \alpha / |L^u|}{V^{u,v} + \alpha} \quad \forall t \geq 2
\]

where \( C(o_t^{u,v}, q_{t-1}^{u,v}) \) is a counting function that returns the number of times that the first activity of the day ends at \( o_t^{u,v} \). Similarly,
As each model outputs $R^2$ and prediction accuracy for each individual, we can plot the distribution of $R^2$ and prediction accuracy for overall performance evaluation. Figure 6 shows the prediction performance for the activity duration and location. Since the first activities are predicted by the initial probability and the remaining activities are predicted by the transition probability, we plot the performance distribution for two types of activities separately. What stands out in the figure is a high degree of individual heterogeneity in terms of predictability. Overall, the IOHMM shows very similar performance as the LSTM model in all prediction tasks. And both IOHMM and LSTM can outperform the first type of baseline models (i.e. LR and MC). This implies that the proposed IOHMM not only has the same predictive capacity as the advanced machine learning model but also has the potential to identify latent activities with model interpretability (details illustrated in Section IV-F).

In terms of the duration prediction (Figure 6a), we observe that IOHMM and LSTM are only slightly better than LR (with mean $R^2 = 0.371, 0.381$, and $0.346$, respectively). This implies that people’s first trip start time on a day has high randomness and is hard to predict as many uncaptured reasons can cause morning departure times to be adjusted. However, for the remaining activities, the IOHMM and LSTM significantly outperform the LR model (with mean $R^2 = 0.692, 0.687$, and $0.563$, respectively).

The results for location prediction (Figure 6b) are similar to those of duration prediction. IOHMM, NG, and LSTM models are slightly better than the MC model in the first activity end location prediction, but significantly better in the prediction of remaining activities. An interesting finding is that, though the duration of the first activity is relatively difficult to predict, the prediction accuracy for the first activity end location (i.e. first trip origin) is high (with a mean of 77.6%, 76.1%, 78.4%, and 74.6% for IOHMM, NG, LSTM, and MC, respectively). This implies that despite randomness in start time, the first trip origins are relatively stable for these frequent public transit users. The location prediction accuracy for the remaining activities is lower than that of the first activities (with a mean of 68.2%, 68.4%, 68.0%, and 51.2% for IOHMM, NG, LSTM, and MC, respectively). This may be attributed to the higher degree of behavioral randomness after leaving home. For the first activity, people are likely to use the nearest rail station around the home. But for remaining activities, people may have more choices that are not easy to capture.

In addition to $R^2$, it is also useful to examine the magnitude of duration prediction errors. The distribution of absolute errors for duration prediction is shown in Figure 7. Overall, errors within 30 minutes account for the highest fraction for all models. For the remaining activities, more than half of the activity duration can be predicted with errors within 1 hour for our IOHMM model. For the first activities, we observe that the LSTM model has a higher density in errors smaller than 30 minutes compared to IOHMM. However, for the remaining activities, IOHMM accounts for a higher density for prediction errors within 30 minutes. As for errors within 1.5 hours, the performance of IOHMM and LSTM models is similar, and both models outperform the LR model. This indicates that LSTM may have more advantages for the first activity duration prediction while IOHMM for remaining activities. This may be because the duration of the first activity (usually staying at home) is hard to predict given the complex interactions of different factors (such as weather, holidays, or even some unobserved factors such as users’ moods). LSTM is more complicated than IOHMM in terms of model structure.
Fig. 6: Prediction performance. “First activities” means the first activity in a day. “Remaining activities” are all other activities in a day excluding the first one. The dash lines represent the mean value.

and the number of parameters. Hence, it may have more prediction power to capture the underlying factor interactions, thus outperforming the first activity duration prediction task.

Fig. 7: Distribution of prediction errors for activity duration. The errors are aggregated in 0.5 hour intervals for better visualization. The solid lines are cumulative density functions (CDF) with colors corresponding to different models.

Figure 8 shows the cumulative distribution of prediction rank for activity end location. The cumulative probability (on the y-axis) at rank $k$ represents the probability that the true activity end location is among the top-$k$ (on the x-axis) most likely outcomes predicted by the model. We observe that, for the first activity, there is more than 90% probability that one of the top 3 predictions in the proposed model is correct. But for the remaining activities, we need to include the top 10 predicted outcomes to achieve 90% probability. The results imply that the origins of the first trips (i.e. first activity end locations) are easier to predict with limited variations than those of following trips. Similarly, IOHMM, NG, and LSTM models achieve comparable (essentially the same) performance in the remaining activity prediction, and both consistently outperform the MC model. However, for the first activities, the NG model becomes worse when counting for more than the top 5 predictions.

Fig. 8: Cumulative distribution of the prediction ranks for activity end location

E. Factors impacting individual mobility predictability

As shown in Figure 6, the prediction performance varies greatly among passengers. Hence, it is worth evaluating which attributes affect the individual’s predictability. We estimated two linear regression models with the $R^2$ (for duration prediction) and prediction accuracy (for location prediction) of IOHMM as dependent variables. Independent variables are factors related to a user’s travel frequency, regularity, fare card type, number of estimated hidden activities, and inferred home location. To reveal how longitudinal behavior changes influence predictability, we introduce the “number of change points” for departure time and visited locations calculated from a Bayesian model in our previous study [46] as new independent variables. These two variables describe the number of substantial behavior pattern changes in terms of departure time and ODs, respectively.

Table II shows the results of estimated coefficients. We observe that the number of days with travel is significantly positive to location prediction, which implies that longer historical trips can increase location predictability. Similarly, the mean number of trips per day is significantly positive for both duration and location prediction. This may be because a high mean number of trips per day reflects longer daily travel sequences, which can potentially make it easier to uncover sequential dependencies and ultimately help with prediction performance. Variables that indicate high travel irregularity, such as the standard deviation (std.) of travel frequency and departure time, number of change points for departure time and locations, have significant negative effects on the prediction performance. We also find that senior passengers’ activity duration is harder to predict. In addition, the activity locations for users living in New Territories (one of the three
main regions of Hong Kong, alongside Hong Kong Island and the Kowloon Peninsula) are easier to predict. This may be because New Territories is further away from the commercial business center of Hong Kong, and its residents generally have less diverse socioeconomic activities outside commuting between home and work. The number of hidden activities has no significant impact on either location or duration prediction accuracy.

### TABLE II: Factors on individual mobility predictability

<table>
<thead>
<tr>
<th>Variables</th>
<th>Duration</th>
<th>End location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.6423 **</td>
<td>0.8441 **</td>
</tr>
<tr>
<td>Total # of days with travel</td>
<td>$3.24 \times 10^{-6}$</td>
<td>$9.92 \times 10^{-5}$ **</td>
</tr>
<tr>
<td>Mean # of trips per day</td>
<td>0.1237 **</td>
<td>0.1074 **</td>
</tr>
<tr>
<td>Std. # of trips per day</td>
<td>-0.1488 **</td>
<td>-0.3058 **</td>
</tr>
<tr>
<td>Std. of departure time of the first trip in a day</td>
<td>-0.0009 **</td>
<td>-0.0005 **</td>
</tr>
<tr>
<td>Student</td>
<td>-0.0298</td>
<td>-0.0094</td>
</tr>
<tr>
<td>Senior</td>
<td>-0.0786 **</td>
<td>0.0023</td>
</tr>
<tr>
<td># change points (time)</td>
<td>-0.0132 **</td>
<td>N.A.</td>
</tr>
<tr>
<td># change points (location)</td>
<td>N.A.</td>
<td>-0.0164 **</td>
</tr>
<tr>
<td># hidden activities ($N^u$)</td>
<td>-0.0014</td>
<td>-0.0033</td>
</tr>
</tbody>
</table>

Number of observations: 500.

Duration $R^2$: 0.295; End location $R^2$: 0.484

* *; $p$-value < 0.01; *; $p$-value < 0.05.

1 Living in Hong Kong Island: 0.0030 0.0103
2 Living in New Territories: 0.0068 0.0165 **
3 Activity start location “Null” (i.e. the one with the highest probability and is much higher than others), and a dominant activity end location “CSW” (Cheung Sha Wan). This is obviously associated with “home” activity because people usually stay at home from the beginning of a day (4:00 AM) to the departure time for working (around 8:00 AM) with a duration of 4 hours. By definition, the first activity of a day is expected, the transition from home to work shows the highest probability. It is worth noting that as the activity after the last trip in a day is omitted from analysis (see Section III-B), the transition from home to work is not revealed in the model. The most likely activity following work is other, which may represent work-based shopping, dining, and entertainment activities (corresponding to results in Section IV-F). And the most likely activity following other is work. This also makes sense because the user usually conducts work-based other activities (such as dining) and after that he/she may need to return to work.

Since the mean activity duration is specified with a linear model in IOHMM (see Section III-D), the estimated parameters in the linear model are useful for understanding
other contextual factors affecting activity duration. Table III summarizes some estimated parameters with interpretability for the same selected individual. As expected, the duration for all activities is higher on rainy days (compared to those without rain). It is worth noting that the sum of three activity duration is not a fixed value, since the activity after the last trip is not considered. Thus, the rainy parameter can be positive for all three activities, which means users may delay their trip departure time when it rains outside. Monday has a positive impact on the work and other activities duration (the impact of Monday on home activity is negligible compared to the other two). Sunday has a negative impact on the work activity duration and a positive impact on home activity duration. On public holidays, the duration of home activities increases and decreases for the other two. Since “other” activities for this individual are usually work-based, all these effects are reasonable.

**TABLE III: Estimated parameters for duration prediction**

<table>
<thead>
<tr>
<th>Activity</th>
<th>Estimated parameters ($\theta_{emr,i}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rainy</td>
</tr>
<tr>
<td>Home</td>
<td>0.159</td>
</tr>
<tr>
<td>Work</td>
<td>0.336</td>
</tr>
<tr>
<td>Others</td>
<td>0.364</td>
</tr>
</tbody>
</table>

**V. CONCLUSION AND DISCUSSION**

This paper proposes an activity-based IOHMM framework to simultaneously predict the time and location of an individual’s next trip using smart card data. The prediction task can be transformed into predicting the hidden activity duration and end location, which enables a natural behavioral representation. Based on a case study with data from Hong Kong’s MTR system, we show that the proposed model has a similar prediction performance as the advanced LSTM model, and significantly outperforms the benchmark models. Unlike LSTM, the proposed activity-based model can also be used to analyze hidden activity patterns, which provides meaningful behavioral interpretation for why an individual makes a certain trip. Therefore, the activity-based prediction framework offers a way to combine the predictive power of machine learning methods and the behavioral interpretability of activity-based models. The estimated activity (or travel purpose) information can facilitate the development of situational awareness in intelligent transportation applications, such as personalized traveler information or targeted travel demand management in public transit systems [47]. Activity-based models have been used extensively in travel demand forecasting and simulation. As demonstrated in [38], IOHMM is well suited to simulating individual activity-travel behavior, which can help public transit agencies to better design policies and plan future services.

A natural extension of this study is to apply the proposed activity-based modeling framework to other data sources and mobility systems. Note that while the Hong Kong MTR system is used as a case study, our approach is agnostic to particular modes. The only information required is the longitudinal observations of individual travel history, including the start/end time, origin, destination, and individual identifier of each trip. Such travel information is generally available in most of the intrinsic mobility data sources. New mobility service providers, such as ride-hailing systems, bike-sharing programs as well as on-demand “pop-up” bus services, also collect individual-level travel records (typically through mobile apps) similar to the transit smart card data, though the predictability of individual mobility may vary by different systems. It is expected that the predictability is higher for public transit systems, because of a higher proportion of commuting trips. A further distinction can be made between stationed systems (e.g., subway, buses, docked bike-sharing) and stationless systems (e.g., taxis, ride-hailing, and dockless bike-sharing) [48]. For the latter, certain spatial aggregation is needed for the proposed method to work.

The proposed methodology is not without limitations. One limitation is that the model requires a long observation period of individual trip records, and does not work well with infrequent or new users with little to no travel histories. Future studies can leverage user clustering techniques to extract similar users’ travel patterns as additional input [49], which can compensate for the sparsity of individual data. Another limitation lies in the assumption of stable travel patterns. However, individual travel patterns may change over time, leading to reduced predictive performance due to domain shift (or distributional shift) issues. To address this, we could potentially adopt a change detection module [46] to guide the update of individual mobility prediction model parameters. One way to do this is through dynamic weighting of data points based on behavior change patterns.

**APPENDIX A**

**SUMMARY OF CONTEXTUAL VARIABLES**

Table IV shows the summary of contextual variables $z_t$. Five different dimensions are considered: weather, day of the week, holidays, last trip information, and historical travel statistics.

**TABLE IV: Summary of contextual variables $z_t$**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rainy (Yes = 1)</td>
<td>D</td>
</tr>
<tr>
<td>Heavy rain (Yes = 1)</td>
<td>D</td>
</tr>
<tr>
<td>Sunny (Yes = 1)</td>
<td>D</td>
</tr>
<tr>
<td>Cloudy (Yes = 1)</td>
<td>D</td>
</tr>
<tr>
<td>Daily mean temperature</td>
<td>C</td>
</tr>
<tr>
<td>Monday - Sunday (Yes = 1)</td>
<td>D</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>National holidays (Yes = 1)</td>
<td>D</td>
</tr>
<tr>
<td>Last activity duration</td>
<td>C</td>
</tr>
<tr>
<td>Last trip travel time</td>
<td>C</td>
</tr>
<tr>
<td># days with travel in past 20 days</td>
<td>I</td>
</tr>
<tr>
<td># consecutive days without travel</td>
<td>I</td>
</tr>
<tr>
<td># trips yesterday</td>
<td>I</td>
</tr>
</tbody>
</table>

Variable type: D: Dummy; C: Continuous; I: Integer.
APPENDIX B

HYPER-PARAMETER SPACE OF THE LSTM MODEL

The hyper-parameters of the LSTM model used in this study (for all individuals) are \( M = 1 \), \( K = 50 \), dropout rate = 0.3, \( l_1 \) regularization = 0, \( l_2 \) regularization = 0, batch size = 30. The model is trained using Adam optimizer (with the default learning rate) with 200 training epochs.

<table>
<thead>
<tr>
<th>Hyper-parameters</th>
<th>Value space</th>
</tr>
</thead>
<tbody>
<tr>
<td># of LSTM layers ( M )</td>
<td>{1, 2, 3, 4, 5}</td>
</tr>
<tr>
<td># of units in LSTM layer ( K )</td>
<td>{30, 50, 100, 150, 200}</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>{0.1, 0.3, 0.5, 0.7}</td>
</tr>
<tr>
<td>( l_1 ) regularization</td>
<td>{0, 10^{-6}, 10^{-4}, 0.01, 0.1, 0.5}</td>
</tr>
<tr>
<td>( l_2 ) regularization</td>
<td>{0, 10^{-6}, 10^{-4}, 0.01, 0.1, 0.5}</td>
</tr>
<tr>
<td>Batch size</td>
<td>{20, 30, 50, 70}</td>
</tr>
</tbody>
</table>

APPENDIX C

ANALYSIS ON NUMBER OF HIDDEN ACTIVITIES

Figure 11 shows the distribution of the number of estimated hidden activities for all 500 samples. Most of the users have three hidden activities. And with the increase in the number of activities, the proportion of users decreases. The high proportion of 3-activity users indicates that most of the frequent users (with at least 300 active days of transit usage during the study period) in the MTR system can be characterized by three major activity patterns: home, work, and other.

Fig. 11: Distribution of number of estimated hidden activities

ACKNOWLEDGMENT

The authors would like to thank Hong Kong’s Mass Transit Railway (MTR) for their support and data availability for this research.

REFERENCES
