Transportation Research Part C 89 (2018) 19–34

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

# Individual mobility prediction using transit smart card data

Zhan Zhao[a], Haris N. Koutsopoulos[b], Jinhua Zhao[c,*]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[b] Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, United States
[c] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

ARTICLE INFO

*Keywords:*
Individual mobility
Next trip prediction
Smart card data
Bayesian *n*-gram model

ABSTRACT

For intelligent urban transportation systems, the ability to predict individual mobility is crucial for personalized traveler information, targeted demand management, and dynamic system operations. Whereas existing methods focus on predicting the next location of users, little is known regarding the prediction of the next trip. The paper develops a methodology for predicting daily individual mobility represented as a chain of trips (including the null set, no travel), each defined as a combination of the trip start time $t$, origin $o$, and destination $d$. To predict individual mobility, we first predict whether the user will travel (*trip making prediction*), and then, if so, predict the attributes of the next trip $(t,o,d)$ (*trip attribute prediction*). Each of the two problems can be further decomposed into two subproblems based on the triggering event. For trip attribute prediction, we propose a new model, based on the Bayesian $n$-gram model used in language modeling, to estimate the probability distribution of the next trip conditional on the previous one. The proposed methodology is tested using the pseudonymized transit smart card records from more than 10,000 users in London, U.K. over two years. Based on regularized logistic regression, our trip making prediction models achieve median accuracy levels of over 80%. The prediction accuracy for trip attributes varies by the attribute considered—around 40% for $t$, 70–80% for $o$ and 60–70% for $d$. Relatively, the first trip of the day is more difficult to predict. Significant variations are found across individuals in terms of the model performance, implying diverse travel behavior patterns.

## 1. Introduction

Individual mobility prediction is a critical enabler for various applications that support intelligent urban transportation systems, such as personalized traveler information, targeted demand management, and dynamic system operations. The success of these applications ultimately leads to enhanced customer experience and improved system performance. The prevalence of personal mobile devices (e.g., mobile phones, smart cards) makes it possible to trace individual digital footprints and potentially discover diverse and complex mobility patterns. However, the problem of predicting individual mobility remains challenging, because travel behavior concerns multiple dimensions (most notably the temporal and spatial dimensions), exhibits longitudinal variability for an individual, and varies across individuals.

Recent years have seen a considerable amount of work dedicated to human mobility modeling based on individual digital traces. The most commonly used data in these studies is mobile phone network data (Pappalardo et al., 2015; Schneider et al., 2013; Song et al., 2010b; Eagle and Pentland, 2009; González et al., 2008). Other data sources, such as GPS data (Zhao et al., 2015), Wi-Fi data

[*] Corresponding author.
*E-mail address:* jinhua@mit.edu (J. Zhao).

https://doi.org/10.1016/j.trc.2018.01.022
Received 13 August 2017; Received in revised form 18 December 2017; Accepted 20 January 2018
Available online 02 February 2018

(Sapiezynski et al., 2015) and social media check-in data (Colombo et al., 2012; Hasan et al., 2013), have also been adopted for mobility studies. This type of data is sensitive and needs to be handled in accordance with privacy legislation. One common trait of these data sources is that they are generated by non-transportation activities, and cannot be interpreted as travel behavior directly. For example, mobile phone network data are generated from cellular network activities including voice calls, text messages, cellular data activities and cell tower handovers. Thus, there is a critical mapping required to bridge the discrepancy between telecommunication behavior and travel behavior. This is not straightforward because the two aspects of behavior are driven by correlated but distinct personal preferences that vary across individuals (Zhao et al., 2016). In this paper, we refer to this type of data as *extrinsic mobility data*.

Extrinsic mobility data capture individual mobility by sampling the user's positions over time. The specific sampling process of a given data source is determined by its data-generating events. Under this data collection mechanism, individual mobility is represented as a series of time-stamped locations, and the prediction problem is framed as that of predicting an individual's next location. Next location prediction is a well studied problem, as will be discussed in Section 2, but it has some limitations in transportation applications. People generally spend most of their time staying at a location, rather than traveling between locations. Thus, over the short term, a user's location will, more often than not, stay the same. Based on the analysis of mobile phone network data, Song et al. (2010b) adopted a representation of individual mobility as a sequence of locations at hourly intervals, and reported a 93% potential predictability of the next location. However, this does not reflect the true predictability of travel behavior. In transportation, the few hours when people travel to a different location are much more important than the majority of hours when they do not. Prediction accuracy of individual locations on an hourly basis tends to be higher when people travel less frequently. For example, Hawelka et al. (2017) found that human mobility is most predictable between 02:00 and 08:00, when most people are asleep. For these reasons, we argue the next location prediction problem is not particularly helpful for transportation applications.

On the other hand, *intrinsic mobility data* are directly collected from urban transportation systems, such as transit smart card data (Goulet-Langlois et al., 2017; Zhong et al., 2015; Hasan et al., 2012) and bike sharing data (Purnama et al., 2015). Unlike extrinsic mobility data, intrinsic mobility data are generated by travel events with each record typically indicating the start or end of a trip. Therefore, intrinsic mobility data provide direct information about individual mobility as a series of trips. The individual mobility prediction problem in this case can be framed as that of predicting an individual's next trip. Unlike time-stamped locations, trips reflect critical travel decision moments, and thus match the actual behavior process of individual mobility. Although some knowledge of people's locations is useful for various location-based services (e.g., recommending a nearby restaurant), trips are more relevant to transportation applications. Intrinsic mobility data are usually mode-specific, and thus the trips captured by such data can be directly mapped to a certain transportation system.

Although a number of algorithms have been proposed for next location prediction, there is no existing model for next trip prediction. The objective of this paper is to define and formulate the specific problems that constitute individual mobility prediction based on intrinsic mobility data, propose a suitable methodology to solve these problems, and examine the predictability of individual mobility across different behavior dimensions using the proposed methodology. The method requires the historic sequence of individual trip records, and is agnostic to any particular mode. In this paper, the rail-based public transportation system in London, U.K. is chosen as a case study and the transit smart card data is the primary data source.

The remainder of the paper is organized as follows. A literature review of the related work on individual mobility prediction is presented in Section 2. Section 3 proposes a new methodological framework, based on Bayesian $n$-gram models, for individual mobility prediction using intrinsic mobility data. Section 4 demonstrates the application of the proposed methodology using pseudonymized transit smart card records of more than 10,000 users over two years. Section 5 concludes the paper with a summary of the main findings, and a discussion of future research directions and potential implications.

## 2. Literature review

A plethora of methods have been proposed to solve the problem of next location prediction. This problem may be formulated in two ways—predicting the location that the user will visit next (Gambs et al., 2012; Mathew et al., 2012; Noulas et al., 2012), or predicting the location of the user in the next time interval (often set as an hour) (Hawelka et al., 2017; Alhasoun et al., 2017; Lu et al., 2013; Calabrese et al., 2010). The main difference is that the former treats individual mobility as a sequence of locations, whereas the latter treats it as a sequence of time intervals with associated locations. As time is a critical dimension of mobility, the latter problem formulation is more valuable for practical applications. However, it requires data with a high data sampling frequency (e.g., the frequency in which new data arrive should be no less than the frequency of the prediction).

The majority of existing methods are based on modeling sequential patterns of individual location histories. One commonly used approach is to model the location sequence of a given user as a Markov Chain (MC). Simple MC models have been shown to be able to achieve high prediction performance (Lu et al., 2013). Gambs et al. (2012) experimented with a previously proposed model called Mobility Markov Chain (MMC) with $k$ previous visited locations. They found that the next location could be predicted with accuracy of 70–95% when $k = 2$, although this did not improve much when $k > 2$. Note that when $k$ increases, the number of transition probabilities grows exponentially, making model estimation increasingly difficult, especially when individual-level data is sparse. As a result, individually fitted MC models are prone to overfitting, and unable to predict locations that users have never visited before. One way to address these issues is to utilize the location histories of other users, in addition to that of the user in focus, for model estimation. The Mixed Markov chain Model (MMM), proposed by Asahara et al. (2011), is an intermediate approach between individual and universal models. MMM predicts the next location by identifying the group to which a particular individual belongs and applying the MC model for that group. Similarly, Mathew et al. (2012) presented a hybrid method of clustering location histories according to their characteristics before training a Hidden Markov Model (HMM) for each cluster. Calabrese et al. (2010) used a weighted combination of individual and collective models for prediction. Alhasoun et al. (2017) proposed a dynamic Bayesian network approach to couple the location sequences of the individuals under consideration with those of their "similar strangers".

The aforementioned methods focus on the next location prediction problem. In this paper, we are interested in predicting the next trip, including its spatial and temporal attributes. This is a problem that has received limited attention. Using GPS data, Gidófalvi and Dong (2012) constructed an inhomogeneous continuous-time Markov model to predict when a user will leave their current location and where they will move next. The model assumes the holding time at any state (or location) is memoryless and exponentially distributed, which is not very realistic. A time-aware language model, T-gram, was presented by Hsieh et al. (2015) to predict human mobility using location check-in data. From data, the T-gram model first extracts a location time distribution (LTD) for each location and a transition time distribution (TTD) for each location pair. Given a sequence of locations with timestamps, it makes a prediction based on how well the sequence matches the learned TTD and LTD. However, LTD and TTD are estimated for each location independently and for all users, and thus the method ignores the sequential dependency between locations and heterogeneity across individuals. These simplifying assumptions will be relaxed in our approach.

There are relatively few studies regarding individual mobility prediction specifically using intrinsic mobility data. Using transit smart card data, Lathia et al. (2013) explored a number of algorithms for personalized prediction of trip duration and demonstrated how prediction accuracy can be improved by incorporating individual behavioral patterns. To predict whether passengers will take the bus in the coming hours/days, a neural network approach was proposed in Dou et al. (2015). Foell et al. (2014) tested several ranking-based approaches to predict the future bus stops accessed by individual passengers, and found collaborative filtering and random walk to be most successful. Although these methods are able to predict a particular aspect of trips, no model has been proposed to predict the next trip as a combination of multiple attributes. In the next section, we present a new method for predicting daily individual mobility as a sequence of trips, each with multiple attributes, based on the previous trip within the day (if exists).

## 3. Methodology

### 3.1. Mobility prediction mechanisms

In this work, we focus on dynamic prediction of individual mobility, meaning that the prediction is updated as new data become available. State-of-the-art methods make predictions on an hourly basis (Song et al., 2010b, Calabrese et al., 2010, Lu et al., 2013). Every hour, for each user, the system predicts the user's location in the next hour. However, this is not well suited to next trip prediction. Time between consecutive trips is relatively long (often multiple hours) and unevenly distributed (large variation). As a result, an alternative prediction mechanism for next trip prediction is proposed, shown in Fig. 1. We split the next trip prediction problem into two subproblems—*trip making prediction* and *trip attribute prediction*. Trip making prediction asks whether the user will travel within the day. Under the condition that the user will travel, we then predict the attributes of the next trip in the task of trip attribute prediction. Both subproblems can be triggered by either a new day or an observation of a new trip for the user. Thus, at the start of each day, the system predicts whether the user will travel within the day. If yes, it then predicts when and where the next trip will be. The prediction is updated every time we observe a new trip, until the end of that day. On the next day, the process is repeated.

The basic prediction interval is set as a day in this case. It serves as a lower bound for prediction frequency, and puts a constraint to what counts as "the next trip". If an individual makes a trip one day and the next one two days later, we disregard the sequential dependency between these two consecutive trips. The day is a natural choice of the basic prediction interval, because it has been identified as one of most fundamental periods of regularity that underlie human mobility patterns (Kim and Kotz, 2007; Eagle and Pentland, 2006). For example, most people consistently start and end their days at the same location. We specifically define that each day spans from 03:00 to 03:00 of the next calendar day, which better matches people's daily activity patterns.

Further distinction can be made between the two triggering events. At the start of a day, a prediction is triggered, without much knowledge about how the user will behave over the period of the day. In contrast, when a prediction is triggered by the observation of

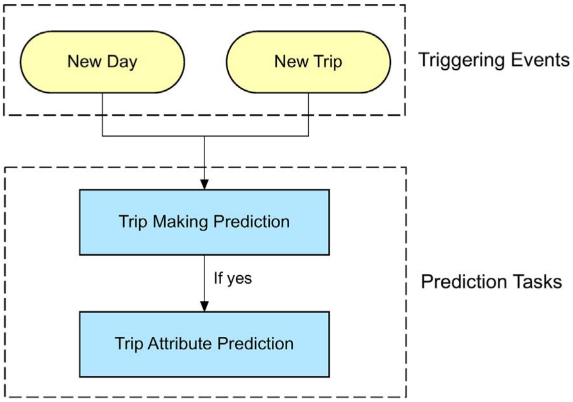

**Fig. 1.** Alternative individual mobility prediction mechanism.

a new trip, the observed trip provides important information for our prediction of the next trip. Whereas the prediction task defines the target to predict, the triggering event determines the choice of features used for prediction.

Based on the proposed prediction mechanism, we define the following four lower-level problems that must be solved for individual mobility prediction:

- **P1A**: at the start of a day, predict whether the user will make a trip within that day;
- **P1B**: based on the last trip observed, predict whether the user will make another trip within that day;
- **P2A**: at the start of a day, predict when and where the first trip will be, given that the user will make a trip within that day (i.e., result of P1A);
- **P2B**: based on the last trip observed, predict when and where the next trip will be, given that the user will make another trip within that day (i.e., result of P1B).

### 3.2. Problem formulation

Each of the four aforementioned problems needs to be solved with a specific model. P1A and P1B are binary classification problems, where only two outcomes are possible. By comparison, P2A and P2B are more challenging, primarily because of the much larger problem space—there are many possible outcomes for each trip attribute, not to mention their combinations. Therefore, in the remainder of this section, we will briefly introduce the formulation of the trip making prediction models, and focus more on the in-depth description of the trip attribute prediction models.

The mobility pattern of an individual can be conceptualized as a sequence of travel events, each of which is characterized by the combination of its attributes (Goulet-Langlois et al., 2017). The specification of these events depends on the problem definition as well as the data characteristics. For individual mobility prediction using intrinsic mobility data, each travel event is set as a trip characterized by three attributes—start time $t$, origin $o$, and destination $d$. We assume all three attributes can be represented as discrete variables, and these variables capture the most fundamental decisions users make when traveling. The need to consider $o$ is because intrinsic mobility data usually only capture a subset of trips pertaining to a certain mode. If complete information of all trips is available, $o$ will always be the same as the $d$ of the last trip.

In this paper, we focus on the individual mobility sequence within a day. Examining the mobility sequence at the daily level, preserves the order of trips in the day, so that we can extract within-day interdependencies between trips. A user's mobility sequence on a given day can be represented as

$$\boldsymbol{S}^{\boldsymbol{u},\boldsymbol{v}} = \{(t_1^{u,v},o_1^{u,v},d_1^{u,v}),(t_2^{u,v},o_2^{u,v},d_2^{u,v}),...,(t_{k^{u,v}}^{u,v},o_{k^{u,v}}^{u,v},d_{k^{u,v}}^{u,v})\} \tag{1}$$

where $\boldsymbol{S}^{\boldsymbol{u},\boldsymbol{v}}$ is the mobility sequence for user $u$ on day $v$, and $k^{u,v}$ is the number of trips made by user $u$ on day $v$. $k^{u,v}$ can be any non-negative integer, but it is unknown before the day ends. In real-time prediction, we do not have to predict $k^{u,v}$ directly. Instead, we incrementally infer whether the next trip exists. Let us use a binary variable $y_i^{u,v}$ to indicate whether the $i$th trip of the day exists, where $i \geq 1$. In trip making prediction, $y_i^{u,v}$ is dynamically estimated after the $(i-1)$th trip is observed. At the start of the day, there are no trips observed, and the goal of P1A is to estimate $y_1^{u,v}$. The goal for P1B is to estimate $y_i^{u,v}$ after the $(i-1)$th trip is observed. In trip attribute prediction, we predict the exact composition of the upcoming trip. The goal of P2A is to estimate $(t_1^{u,v},o_1^{u,v},d_1^{u,v})$ at the start of the day, and the goal of P2B is to estimate $(t_i^{u,v},o_i^{u,v},d_i^{u,v})$ after the $(i-1)$th trip.

For simplicity, superscripts $u$ and $v$ are omitted and all remaining notation is defined with respect to a given individual on a given day. Let us assume that the daily travel sequence shown in Eq. (1) unfolds as a Markov chain; a trip only depends on the previous trip of the day (if exists). With no previous trip available, the first trip of the day may be predicted based on the characteristics of the day (e.g., the day of week), denoted as $\boldsymbol{F}$. Table 1 summarizes the target probability distribution to be estimated for each of the four problems.

As $y_i$ is a binary variable, P1A and P1B can be solved with a logistic regression model. Specifically, "L2" regularization is used to penalize model complexity and avoid overfitting (Ng, 2004). The difference between P1A and P1B is that there likely exists autocorrelation between consecutive days in P1A. This is accounted for by including lagged dependent variables as part of explanatory variables $\boldsymbol{F}$, as in an autoregressive (AR) model.

For P2A and P2B, it is not easy to directly compute the joint probability of $P(t_1,o_1,d_1|\boldsymbol{F})$ and especially $P(t_i,o_i,d_i|t_{i-1},o_{i-1},d_{i-1})$, as there are a large number of possible combinations and only limited amount of observations at the individual level. Therefore, a new method is proposed in Section 3.3.

**Table 1**
Problem formulation.

| Problem | Target to Estimate |
|---|---|
| P1A | $P(y_1\|\boldsymbol{F})$ |
| P1B | $P(y_i\|t_{i-1},o_{i-1},d_{i-1})$ |
| P2A | $P(t_1,o_1,d_1\|\boldsymbol{F})$ |
| P2B | $P(t_i,o_i,d_i\|t_{i-1},o_{i-1},d_{i-1})$ |

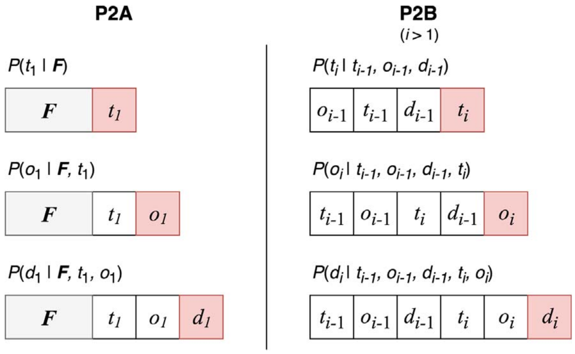

**Fig. 2.** Illustration of the mobility n-gram model as a combination for n-gram models.

### 3.3. Mobility n-gram model

Using the chain rule, the joint probability can be factorized into the product of three conditional probabilities, which are easier to compute separately. For P2A and P2B, the target joint probability can be rewritten as

$$P(t_1,o_1,d_1|\boldsymbol{F}) = P(t_1|\boldsymbol{F})\cdot P(o_i|\boldsymbol{F},t_1)\cdot P(d_1|\boldsymbol{F},t_1,o_1) \tag{2}$$

$$P(t_i,o_i,d_i|t_{i-1},o_{i-1},d_{i-1}) = P(t_i|t_{i-1},o_{i-1},d_{i-1})\cdot P(o_i|t_{i-1},o_{i-1},d_{i-1},t_i)\cdot P(d_i|t_{i-1},o_{i-1},d_{i-1},t_i,o_i) \tag{3}$$

If we assume all variables are discrete, the components in Eqs. (2) and (3) can be individually estimated using variations of *n*-gram models. In natural language processing, *n*-gram models are commonly used to estimate the probability of different word sequences, which supports a range of applications such as speech recognition and machine translation. An *n*-gram model predicts the next word in a sequence in the form of an $O(n-1)$ Markov model (Jurafsky and Martin, 2008). In other words, it assumes that the probability of a word only depends on the previous $(n-1)$ words. In this paper, we propose a mobility *n*-gram model for trip attribute prediction, in which an *n*-gram model is built for each component in Eqs. (2) and (3), as illustrated in Fig. 2. The last element in red background is the target variable to be predicted, and the elements to its left can be regarded as the "context" used for prediction.

Because of the large number of words and their combinations, *n*-gram models also face the sparse data problem. Some perfectly acceptable phrases may not be observed from a training corpus, but it is possible that they appear in a new document. The Maximum Likelihood Estimation (MLE) method would assign a zero probability to any pattern unseen in the training data. Furthermore, it also produces poor results when the number of observations of a certain pattern is non-zero but still small (Jurafsky and Martin, 2008). For these reasons, smoothing is used to assign a non-zero possibility mass to any possible unseen pattern. Existing smoothing methods for language modeling range from simple Laplace smoothing to more sophisticated techniques such as Good-Turing discounting, Katz's back-off model or Kneser-Ney smoothing (Chen and Goodman, 1999). Although these methods have shown to perform well in practice, most of them take a rather ad hoc approach, making it difficult to apply them beyond language modeling. Bayesian *n*-gram models, however, have explicitly declared assumptions and are easy to improve by incorporating additional domain knowledge (Teh, 2006). MacKay and Peto (1994) introduced a Bayesian *n*-gram model with Dirichlet prior, which is what our model is based on.

Let us use $(x_{1:n-1},x_n)$ to denote an *n*-gram, where $x_{1:n-1}$ and $x_n$ represent the context and target variables of the *n*-gram, respectively. When making a prediction given context $x_{1:n-1}$, the target variable $x_n$ is unknown, and we estimate the probability of $x_n$ based on the following equation:

$$\hat{P}(x_n|x_{1:n-1}) = \frac{C(x_{1:n-1},x_n) + \alpha m_{x_n|x_{1:n-1}}}{\sum_{x_n'\in V} C(x_{1:n-1},x_n') + \alpha},(n>1) \tag{4}$$

where $C(x_{1:n-1},x_n)$ is the time-smoothed counting function that returns the estimated frequency of $(x_{1:n-1},x_n)$ in the travel history of the individual (described in Appendix A), $\boldsymbol{V}$ is the set of values that $x_n$ is drawn from, $\alpha$ is a scalar concentration parameter of the Dirichlet prior, and $m_{x_n|x_{1:n-1}}$ is the expectation of the prior probability of $P(x_n|x_{1:n-1})$.

Because of data sparsity, $\sum_{x_n'\in V} C(x_{1:n-1},x_n')$ in Eq. (4) is often small, and as a result the choice of the expected prior probability $m_{x_n|x_{1:n-1}}$ significantly influences model performance. In our model, $m_{x_n|x_{1:n-1}}$ is assumed to be a linear combination of probability estimates from a couple of alternative models—an individual $(n-1)$-gram model and a population *n*-gram model. The two probability are balanced by a parameter $\beta \in [0,1]$, as shown in Eq. (5).

$$\hat{m}_{x_n|x_{1:n-1}} = \beta\hat{P}(x_n|x_{2:n-1}) + (1-\beta)\hat{P}_0(x_n|x_{1:n-1}) \tag{5}$$

$\hat{P}(x_n|x_{2:n-1})$ is used as a *back-off prior* for $\hat{P}(x_n|x_{1:n-1})$, which is named after Katz's back-off model (Katz, 1987). The back-off prior is itself estimated from a simpler $(n-1)$-gram model with less specific context (i.e., one less context variable $x_1$) but more reliable estimations (i.e., more observations of the context $x_{2:n-1}$). It helps prevent overfitting when the frequency of the original context $x_{1:n-1}$ is low. Each back-off prior has its own back-off prior that is recursively calculated in a similar fashion as shown in Eq. (5); the back-off prior of $\hat{P}(x_n|x_{2:n-1})$ is $\hat{P}(x_n|x_{3:n-1})$, and so on. $\hat{P}_0(x_n|x_{1:n-1})$ is used as a *collective prior*, aiming to improve the prediction by crowdsourcing the behavior of other users in the same context $x_{1:n-1}$. $\hat{P}_0(x_n|x_{1:n-1})$ is estimated using Eq. (6), which relies on its own back-off prior. The inclusion of the collective prior allows the model to be applied to new users with no travel histories.

$$\widehat{P}_0(x_n|x_{1:n-1}) = \frac{C_0(x_{1:n-1},x_n) + \alpha_0\widehat{P}_0(x_n|x_{2:n-1})}{\sum_{x_n'\in \boldsymbol{V}} C_0(x_{1:n-1},x_n') + \alpha_0} \tag{6}$$

where $C_0(x_{1:n-1},x_n)$ is a time-smoothed counting function based on the travel histories of all users, and $\alpha_0$ is the concentration parameter for the population model.

Because of the use of back-off priors, the configuration of the context is important. Generally, context variables with a stronger interdependency with the target variable should be placed closer to the target variable (i.e., to the right). Such interdependency may be measured with mutual information, a metric that quantifies the amount of information obtained about one random variable through another. In machine learning, it is often used for feature selection (Peng et al., 2005). The context configuration shown in Fig. 2 is based on both the analysis of mutual information (see Appendix B) and the preliminary evaluation of prediction accuracies.

The back-off priors are computed recursively, until the number of the context variables reaches 0. In such cases, the following equations are used to estimate the probability of $x_n$:

$$\widehat{P}(x_n) = \frac{C(x_n) + \alpha\widehat{P}_0(x_n)}{\sum_{x_n'\in \boldsymbol{V}} C(x_n') + \alpha} \tag{7}$$

$$\widehat{P}_0(x_n) = \frac{C_0(x_n) + \alpha_0\frac{1}{|\boldsymbol{V}|}}{\sum_{x_n'\in \boldsymbol{V}} C_0(x_n') + \alpha_0} \tag{8}$$

Note that Eq. (8) is smoothed by a uniform prior, so that Eq. (4) can return a positive probability for any possible $x_n$.

For better model performance, parameters $\alpha$ and $\beta$ should be set individual-specific. To automate the parameter selection process, a gradient ascent procedure is implemented to search for the approximately optimal $\alpha$ and $\beta$ based on likelihood maximization. Specifically, during the training stage, a subset of the training data is held out for likelihood estimation, and the values of $\alpha$ and $\beta$ are updated so that the likelihood of this subset under the model is maximized.

### 3.4. Model evaluation

For both trip making prediction (i.e., P1A and P1B) and trip attribute prediction (i.e., P2A and P2B), a model is trained and tested for each user. Prediction accuracy is used as the main model performance metric, which is defined as the percentage of the test cases where the predicted label matches the true label. In addition, cross entropy, also known as log loss, is used to evaluate probabilistic predictions. It is defined as the average negative log-likelihood of the true labels given a probabilistic model's predictions in the sample. A lower cross entropy indicates better model performance (Jurafsky and Martin, 2008). For binary classification models, the F1 score is a commonly used performance metric, and computed as a weighted average of the precision and recall. Like prediction accuracy, it ranges from 0 to 1, with 1 being the best. The F1 score is only used for evaluation of trip making prediction, because it is not well defined for multi-class classification problems. As the models are developed and evaluated at the individual level, the median of the model performance metrics is used as an aggregate performance metric. The median is chosen because it illustrates how the models perform on an "average" user without being skewed by a few "outlier" users in the sample.

For the baseline model to compare against for the case of trip making prediction, we use a logistic regression model with constants only. Such a model is equivalent to setting the probabilities equal to the actual shares in the training set.

There is no existing method for trip attribute prediction, and we develop a Markov model for performance benchmarking. It has been shown in the literature that a first-order Markov Chain, or MC(1), can approach the limit of predictability for the next location prediction problem, and increasing the order does not necessarily improve the prediction performance (Lu et al., 2013). Our model consists of one MC(1) in the temporal dimension and another MC(1) in the spatial dimension. In the temporal MC(1), $t_i$ is only dependent on $t_{i-1}$; in the spatial MC(1), $o_i$ is only dependent on $d_{i-1}$, and $d_i$ on $o_i$. We will refer to this model as 2-MC(1) in the remainder of the paper. It can be regarded as a simplified version of the mobility $n$-gram model.

Let us assume that the set of possible time values is $\boldsymbol{T}$, the set for the location values is $\boldsymbol{L}$, and the total number of days in which a trip takes place in the training data for a given user is $M$. Then the initial distributions of the 2-MC(1) model are calculated as follows:

$$P(t_1) = \frac{c(t_1) + \alpha/|\boldsymbol{T}|}{M + \alpha} \tag{9}$$

$$P(o_1) = \frac{c(o_1) + \alpha/|\boldsymbol{L}|}{M + \alpha} \tag{10}$$

where $c(t_1)$, or $c(o_1)$, denotes a counting function that returns the number of times the first trip of the day starts at $t_1$, or from origin $o_1$, in the training data. Similar to Eq. (8), the parameter $\alpha$ is used for smoothing so that a non-zero probability is generated for any possible value.

The transition probabilities of 2-MC(1) are calculated as follows:

$$P(t_i|t_{i-1}) = \frac{c(t_{i-1},t_i) + \alpha/|\boldsymbol{T}|}{\sum_{t_i'\in \boldsymbol{T}} c(t_{i-1},t_i') + \alpha}, \quad \text{where } i \geqslant 2 \tag{11}$$

$$P(o_i|d_{i-1}) = \frac{c(d_{i-1},o_i) + \alpha/|\boldsymbol{L}|}{\sum_{o_i' \in \boldsymbol{L}} c(d_{i-1},o_i') + \alpha}, \quad \text{where } i \geqslant 2 \tag{12}$$

$$P(d_i|o_i) = \frac{c(o_i,d_i) + \alpha/|\boldsymbol{L}|}{\sum_{d_i' \in \boldsymbol{L}} c(o_i,d_i') + \alpha}, \text{where } i \geqslant 1 \tag{13}$$

where $c(t_{i-1},t_i)$ denotes a counting function that returns the number of times a trip starting at $t_{i-1}$ is followed (in the same day) by another trip starting at $t_i$, $c(d_{i-1},o_i)$ returns the number of times a trip ending at $d_{i-1}$ is followed (in the same day) by another trip starting from $o_i$, and $c(o_i,d_i)$ returns the number of times a trip goes from $o_i$ to $d_i$.

## 4. Results

### 4.1. Data

The specific dataset used for the analysis contains pseudonymized transit smart card records from 10,479 anonymous users between September 2014 and September 2016 in London, U.K. Note that a user may possess multiple cards, and a card may be shared by multiple users. However, it is difficult to identify these incidences. In this paper, we assume each smart card corresponds to one and only one user. The public transportation system in London consists of several modes. Due to data availability, the scope of the analysis in this paper is limited to the rail-based systems, including London Underground, Overground, and part of the National Rail. Preprocessing was conducted to select users with at least 60 active days of transit usage during the study period, which excludes occasional users and short-term visitors such as tourists. The mobility prediction for infrequent users and short-term visitors is an area that requires further research. To achieve reasonable prediction performance, a minimum amount of personal travel history is typically required. During the study period, these users generated 3,832,491 trips over 1,921,503 active user-days, resulting in an average trip rate of 1.99 trips per active day. Overall, 715 distinct stations were accessed by these users. We assume that this is the complete spatial choice set shared by all users, meaning any user has a non-zero probability of traveling to any of the 715 stations. Given the nature of the data, the findings presented in this paper are most relevant to public transportation users.

### 4.2. Mobility patterns

A preliminary exploration of the dataset reveals several basic mobility patterns. Fig. 3(a) shows that the number of stations accessed by transit users varies considerably across individuals. Some users access more than 100 stations over the two-year period, and are likely to use public transportation for a range of trip purposes, whereas those who only access a small number of stations probably ride trains exclusively for commuting. The distribution of the number of trips per day is shown in Fig. 3(b). Apparently, a user typically makes two trips a day. This distribution is particularly relevant for P1B; the number of trips observed so far is an important predictor for whether the user will make another trip within the day. Note that only rail-based trips are considered in the case study. Thus, this distribution is an underrepresentation of the true travel intensity of users.

It is known that the occurrences of words in a given natural language corpus follow Zipf's law, meaning the frequency $f_k$ of the $k$th most frequent word is approximately proportional to $k^{-\zeta}$, where $\zeta$ is a nonnegative constant value characterizing the shape of the distribution. Although it has been shown that the location visitation frequency distribution of an individual conforms to Zipf's law (González et al., 2008; Song et al., 2010a), we find that individual temporal choices also follow a similar pattern. Fig. 3(c) shows the rank-frequency distribution of trip attributes on a log-log scale. The rank (on the x-axis) is determined based on the frequency of occurrences at the individual level, and the probability (on the y-axis) indicates the normalized frequencies for the top-ranked values of $t$,$o$, and $d$. As each individual has different probabilities, only the median values are shown in the figure. Based on the slopes of the distributions, it is apparent that $\zeta$ is larger for spatial choices than for temporal choices, suggesting choices of $o$ and $d$ tend to be more concentrated to a few locations. To understand the impact of temporal resolutions, $t$ is aggregated to different levels, i.e., 15 min, 30 min, and 1 h. Though the specific value of $\zeta$ decreases with the resolution, the general shape of the distribution does not vary significantly. In this paper, we aggregate time to hour-long bands, without loss of generality.

To measure the variation of individual travel behavior, we calculate the entropy of $t$, $o$, and $d$. Higher entropy means the variable has higher uncertainty, and is generally more difficult to predict (Song et al., 2010b). In Fig. 3(d), the colored bars show the median entropy for each trip attribute, with the black line segments indicating the interquartile range, i.e., the range between the 25-percentile and 75-percentile values. We find that $t$ (aggregated to hours), albeit with lowest cardinality, has the highest entropy on average, implying high degree of variability in terms of individual temporal choices. Overall, $o$ and $d$ have very similar entropy values, with $d$ slightly higher than $o$. If we have the complete travel records of these users, $o$ and $d$ should have the exact same entropy. One possible interpretation is that people are more likely to use alternative modes (e.g., taxi) to return home at late night or early morning when the public transportation service is relatively limited and walking to train stations is perceived to be less safe. As the first trip of the day is treated separately from the rest in our mobility prediction problem, the entropy of these two subgroups of trips are compared. It reveals that the first trip of the day tends to have much lower entropy of $o$. This is not surprising because the first trip of the day usually starts from home. In contrast, the trips other than the first trip of the day exhibits lower entropy of $d$, because a large portion of these are afternoon commuting trips ending at home.

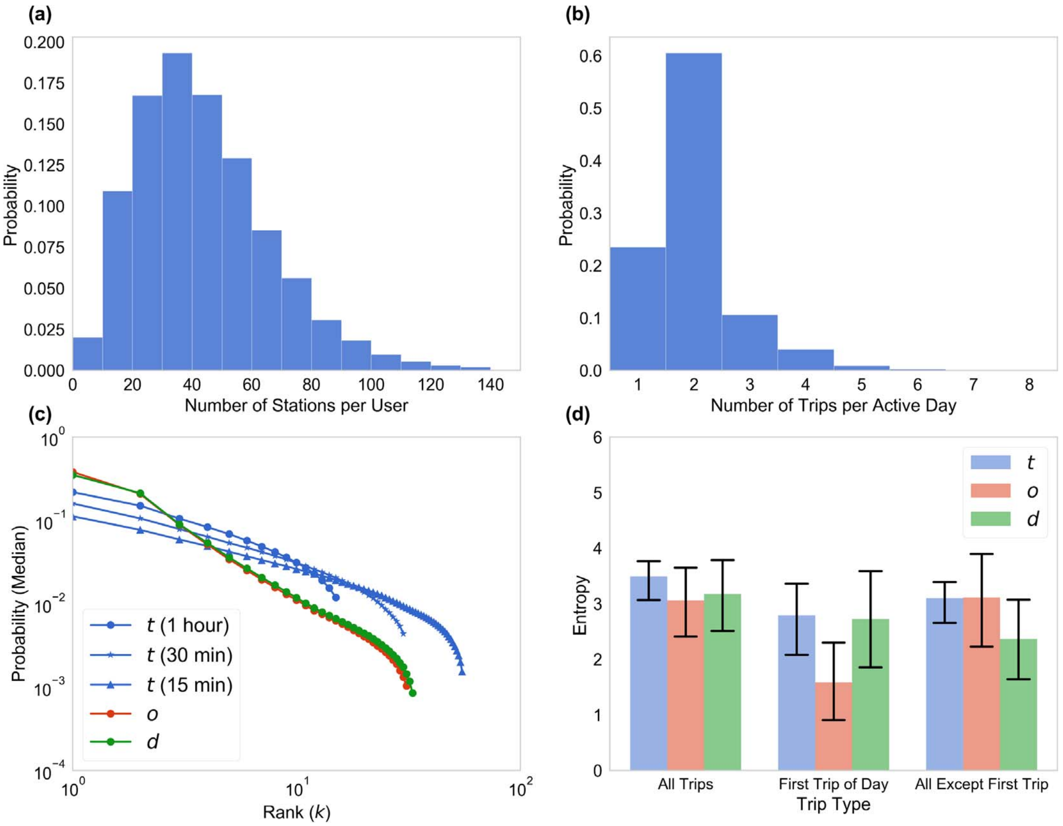

**Fig. 3.** Distribution of individual travel decisions.

### 4.3. Trip making prediction

For P1A, we make predictions based on the characteristics of the day as well as the user's behavior in previous days. A set of features (i.e., $\boldsymbol{F}$) are extracted for a given day of a given user. These features and their estimated coefficients are summarized in Table 2. Note that coefficients are estimated for each individual, and are mostly significant. Only the median and interquartile range (in parentheses) are shown in the table. In general, users are more likely to travel on weekdays and less on weekends and holidays, as expected. We find that the behavior of travel or not on a day exhibits streakiness. Users are more likely to travel, if they do the previous day, or have a higher travel frequency in recent days. On the other hand, they are less likely to travel when there are a long streak of days with no trip. For these coefficients, the interquartile range is larger, in some cases significantly larger, than the median, evidencing high degree of behavioral heterogeneity across individuals.

For P1B, the explanatory variables include the characteristics of the last observed trip, including the trip start time $t_{i-1}$, origin $o_{i-1}$, destination $d_{i-1}$, as well as the order of the next trip within the day $i$. $d_{i-1}$ is expected to have a great influence because the last trip of the day usually ends at the same location (most likely home). In addition, increase in $t_{i-1}$ and $i$ is likely to decrease the probability of the user making another trip. In the model, all variables are treated as categorical variables, converted to a series of binary variables

**Table 2**
List of prediction features for P1A.

| Feature | Explanation | Estimate |
|---|---|---|
| Monday | If the day is a Monday | 0.40 (1.10) |
| Tuesday | If the day is a Tuesday | 0.28 (0.80) |
| Wednesday | If the day is a Wednesday | 0.22 (0.72) |
| Thursday | If the day is a Thursday | 0.18 (0.65) |
| Friday | If the day is a Friday | 0.18 (0.61) |
| Saturday | If the day is a Saturday | −0.85 (1.47) |
| Sunday | If the day is a Sunday | −1.17 (1.26) |
| Holiday | If the day is a bank holiday | −0.83 (0.98) |
| Previous day | If the user traveled in the previous day | 0.95 (1.01) |
| Travel frequency | Number of days with trips in the past 20 days | 0.75 (1.51) |
| Non-travel days | Number of consecutive days with no travel prior to the day | −0.06 (0.12) |

**Table 3**
Performance metrics (median) for trip making prediction.

| | P1A | | P1B | |
|---|---|---|---|---|
| | Logistic Reg. | Naive Model | Logistic Reg. | Naive Model |
| Prediction Accuracy | 80.4% | 57.8% | 86.5% | 51.2% |
| F1 Score | 0.723 | 0.407 | 0.865 | 0.490 |
| Cross Entropy | 0.676 | 0.890 | 0.491 | 0.994 |

in implementation. As different individuals travel to different locations at different times, they have different combinations of binary variables. The specific coefficients are not presented here.

To evaluate the model performance, 5-fold cross validation is used. Specifically, the original sample is randomly partitioned into 5 equal sized subsamples. At each iteration, a single subsample is retained as the test set, and the remaining 4 subsamples are used as training set. The process is then repeated 5 times, with each of the 5 subsamples used exactly once as the validation data. The results from the 5 iterations are then averaged to produce a single estimation. The resulting prediction performance measures for P1A and P1B are summarized in Table 3. As expected, the proposed models for P1A and P1B significantly outperform the naive model. The prediction performance is higher for P1B. The distribution of the trip making prediction accuracy over users is shown in Fig. 4. P1A and P1B both display some interpersonal variability. The median of the distribution is marked with a dashed vertical line. Although the distribution for P1B has a higher median value, its variation across users is slightly larger than P1A.

### 4.4. Trip attribute prediction

For trip attribute prediction, we partition the personal daily mobility sequences of each user into training and test sets. The test set consists of the sequences from 30 randomly selected active days. The remaining sequences form the training set. A mobility *n*-gram model is developed for each user based on their own training data. For daily features $\boldsymbol{F}$, we only considers the day of the week as the primary factor affecting $t_1$, $o_1$, and $d_1$.

In model implementation, we consider two scenarios—*sequential prediction* and *simultaneous prediction*. Sequential prediction assumes $t_i$, $o_i$, and $d_i$ can be predicted in a sequential manner, where $o_i$ is predicted based on the true $t_i$, and $d_i$ is predicted based on the true $t_i$ and $o_i$. In this way, we can decompose the mobility *n*-gram model into three separate *n*-gram models and assess their performance individually. In simultaneous prediction, we need to simultaneously predict the combination of all these attributes $(t_i,o_i,d_i)$. To do this, we need to predict $o_i$ without knowing $t_i$, and $d_i$ without knowing $t_i$ and $o_i$. Whereas simultaneous prediction is more realistic, sequential prediction provides a better way to compare predictability of different trip attributes and diagnose the model.

### 4.4.1. Sequential prediction

The sequential prediction performance measures for P2A and P2B are summarized in Table 4. For both problems, the mobility *n*-gram model outperforms the 2-MC(1) model, but the margin is much less significant for P2A. Note that 2-MC(1) does not accounts for

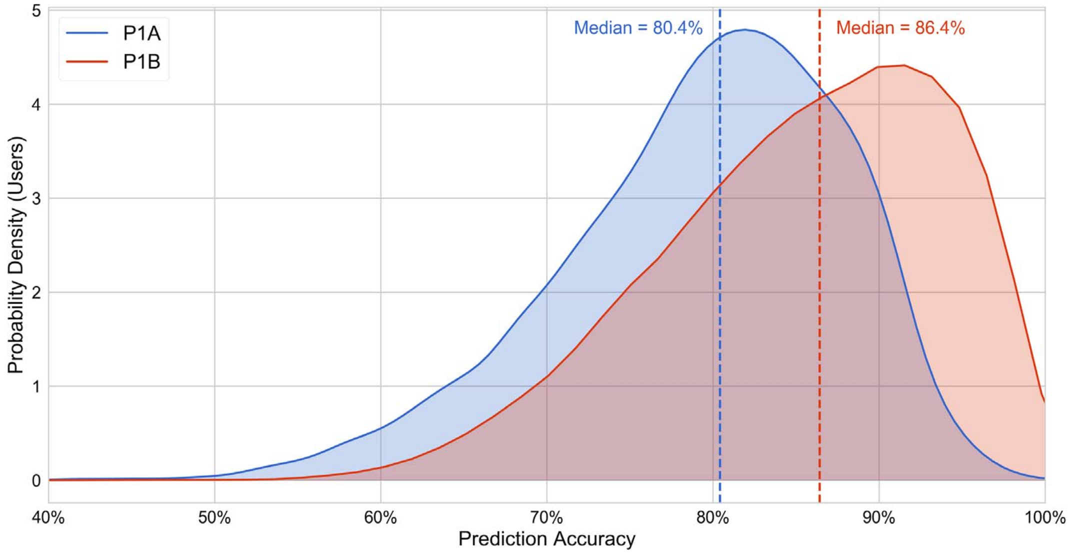

**Fig. 4.** Trip making prediction accuracy.

**Table 4**
Performance metrics (median) for sequential prediction of trip attributes.

| | P2A | | P2B | |
|---|---|---|---|---|
| | *N*-Gram | 2-MC(1) | *N*-Gram | 2-MC(1) |
| Prediction Accuracy | | | | |
| *t* | 40.0% | 36.7% | 41.9% | 32.0% |
| *o* | 73.3% | 66.7% | 81.1% | 70.9% |
| *d* | 63.3% | 60.0% | 70.3% | 57.9% |
| Cross Entropy | | | | |
| *t* | 2.85 | 3.22 | 2.67 | 4.19 |
| *o* | 1.84 | 2.01 | 1.37 | 2.49 |
| *d* | 3.24 | 3.26 | 2.53 | 3.64 |

the day of the week effect. This implies that the day of the week is a limited predictor for the first trip of the day. However, once we have observed a trip, the predictability of the following trips within the day increases, as evidenced by the higher performance for P2B than P2A across all trip attributes.

$o_i$, with the highest prediction accuracy and lowest cross entropy, is clearly the most predictable trip attribute. This is to be expected, because $o_i$ should be close to (and often the same as) $d_{i-1}$. Overall, $t_i$ is the most difficult attribute to predict. For an average user, the start hour of the next trip can only be predicted around 40% of the time, significantly lower than the prediction accuracy of $o_i$ and $d_i$. It suggests that people's temporal choices may be more flexible than their spatial choices. For example, a commuter visits the same location in the morning repeatedly (e.g., going to work), but the exact time of the trip may vary. Although the prediction accuracy for $d_i$ is significantly higher than that of $t_i$, their median cross entropy measures are similar. It indicates that the time prediction model assigns relatively high probability to the true label even if the most likely prediction is incorrect. Because of the relatively low dimensionality of $t_i$, on average each label is assigned a higher probability. As the aggregated hourly intervals are naturally ordered (0–23), we can calculate the magnitude of errors for time predictions, as described in Appendix C. It is found the distribution of the prediction errors centers around 0, meaning when time predictions are wrong they tend to be relatively closer to the true time.

The distributions of sequential prediction accuracies over users are shown in Fig. 5, where the dashed vertical lines indicate the medians of the distributions. Overall variation across users is significant, implying a high degree of individual heterogeneity. Such heterogeneity is particularly pronounced for $d_i$, which has the smallest mode. Compared to P2A, it is apparent that P2B has less performance variation. Furthermore, the inter-personal variability in prediction accuracy is interdependent across trip attributes; it is found that a user with high predictability in one aspect of human mobility tends to exhibit high predictability in other aspects as well (see Appendix D).

Another way to assess a predictive model is to examine the prediction ranking. All possible outcomes can be ranked based on the probabilities estimated by the model; the outcome with the highest probability is ranked first, and so on. A robust model should on

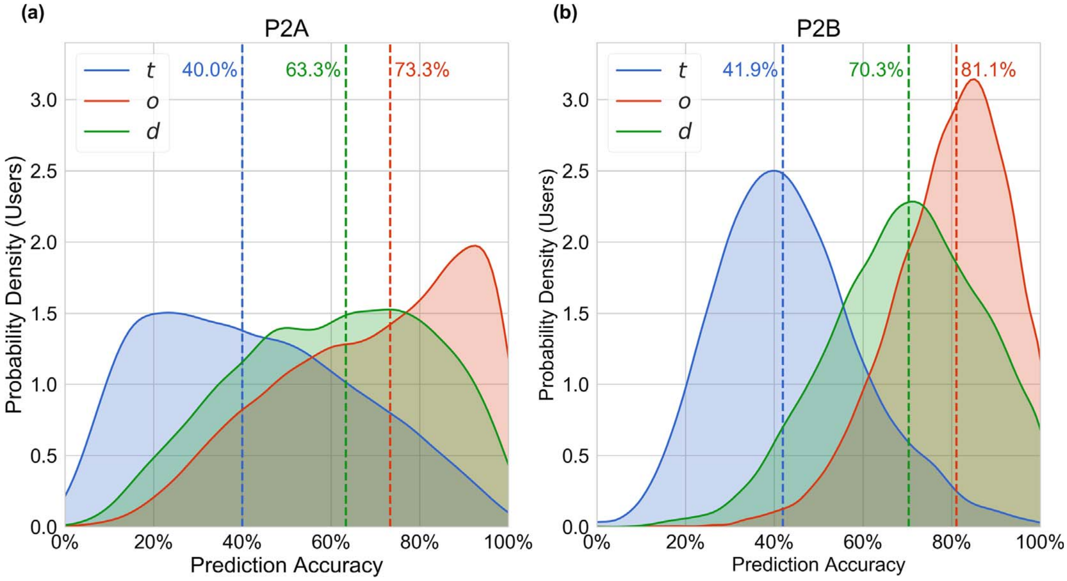

**Fig. 5.** Sequential prediction accuracy of trip attributes.

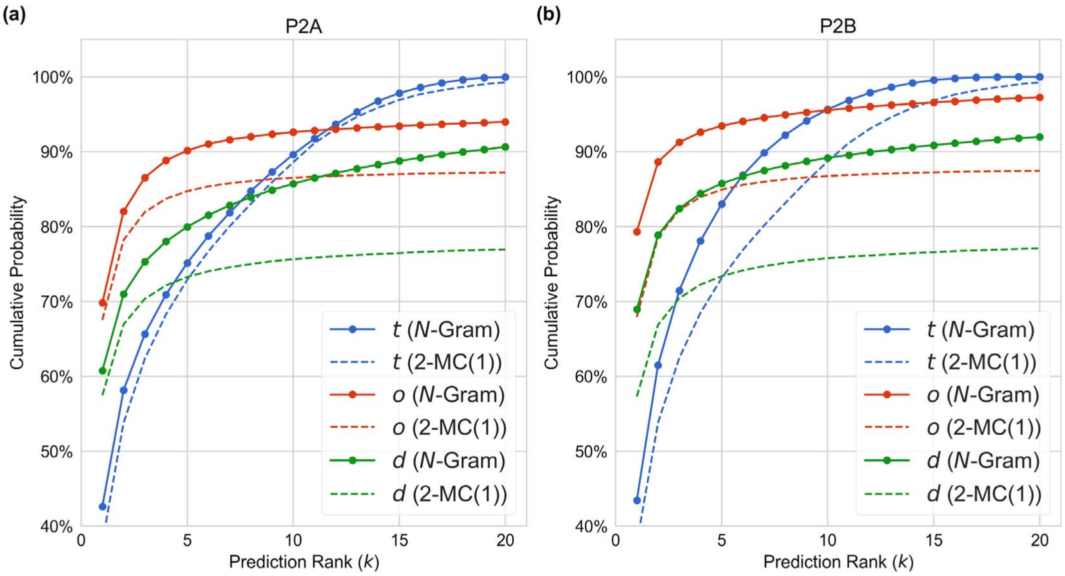

**Fig. 6.** Cumulative distribution of the prediction ranks for trip attribute prediction.

average assign a higher rank (closer to 1) to the true outcome. In Fig. 6 we show the cumulative distribution of the ranking of the true trip attribute. The cumulative probability (on the y-axis) indicates the probability that the true outcome is among the top-$k$ ($k$ = 1,2,...,20, on the x-axis) most likely outcomes predicted by the model. We find that, for prediction of $t_i$ and $o_i$, there is a probability of over 90% that one of the top 10 predictions is correct. For $d_i$, this probability is lower but still over 85%. Except for $t_1$, the proposed model is consistently better than 2-MC(1), especially for P2B. Note that as the rank increases, the cumulative probability approaches 1 much faster for $t_i$ than for $o_i$ and $d_i$. Again, this is due to the smaller cardinality of time. As there are only 24 hourly bands, if we make 20 most likely predictions of $t_i$, it is almost certain one of them is correct. The same thing cannot be said for $o_i$ and $d_i$ as their cardinality is much higher than 20. For some applications, *false negatives* are more costly than *false positives*. For example, if there is a potentially dangerous incident occurring at a certain location, failing to alert a relevant user would have a much more serious consequence than sending a false alarm. In this case, it is useful to notify all users for whom the said location is predicted to be among the likely places they will visit in the near future. In such cases, we may consider more than one most likely outcomes, and Fig. 6 can help us decide the number of outcomes to consider.

### 4.4.2. Simultaneous prediction

Because of the probabilistic nature of the mobility $n$-gram model, it can be extended to predict multiple variables together. For simultaneous prediction, we first estimate the conditional probability distributions of $t_i$, $o_i$, and $d_i$ separately, and then use the chain rule to take the product of the three to obtain the joint probability distribution. For implementation purposes, we do not have to traverse all possible combinations of $(t_i,o_i,d_i)$, which would be computationally expensive. Instead, at each stage, we only consider the 10 most likely outcomes. Specifically, for trip attribute prediction, we first predict the 10 most likely values of $t_i$; for each $t_i$, we then predict the 10 most likely values of $o_i$; for each pair of $(t_i,o_i)$, we predict the 10 most likely values of $d_i$. Finally, we obtain 1000 possible candidate combinations of $(t_i,o_i,d_i)$ and their associated probabilities. The combination associated with the highest probability is selected as the predicted trip.

The simultaneous prediction performance is summarized in Table 5. A prediction is considered to be correct if and only if the

**Table 5**
Performance metrics (median) for simultaneous prediction of trip attributes.

| | P2A | | P2B | |
|---|---|---|---|---|
| | *N*-Gram | 2-MC(1) | *N*-Gram | 2-MC(1) |
| Prediction Accuracy | | | | |
| $(t,o,d)$ | 23.3% | 23.3% | 29.4% | 17.6% |
| $t$ | 40.0% | 36.7% | 42.1% | 31.5% |
| $o$ | 66.7% | 66.7% | 78.3% | 70.8% |
| $d$ | 46.7% | 46.7% | 65.5% | 55.3% |
| Cross Entropy | | | | |
| $(t,o,d)$ | 8.01 | 8.55 | 6.74 | 10.54 |

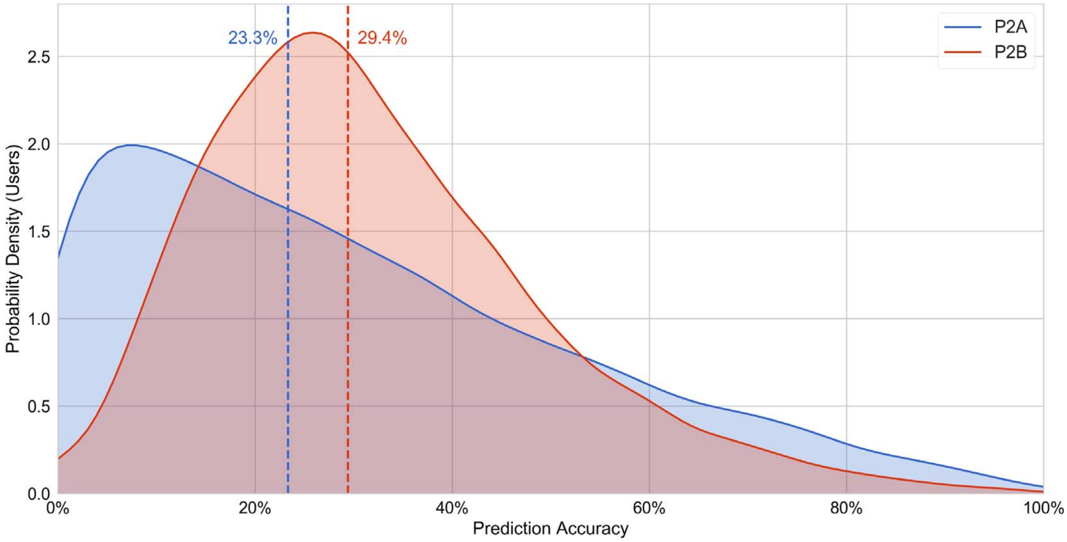

**Fig. 7.** Simultaneous prediction accuracy of trip attribute combinations.

predicted trip attribute combination matches the true combination in every attribute. The median prediction accuracy is 23.3% for P2A, and 29.4% for P2B. Similarly, the median cross entropy is lower for P2B than P2A. Although the mobility *n*-gram model only marginally outperforms 2-MC(1) (in terms of the cross entropy) for P2A, it is significantly better for P2B. Note that there are a limited number (i.e., 31) of possible prediction accuracy results for P2A, because each user has 30 trips to predict in the test set. Two models with the same median prediction accuracy does not mean that they are exactly the same. Fig. 7 shows that the distribution of the prediction accuracy for the trip attribute combination ($t_i$,$o_i$,$d_i$). Again, a high degree of individual heterogeneity is demonstrated. Interestingly, there are a larger number of users with a prediction accuracy of 60% or above for P2A than P2B, despite P2A has a lower median value. Although the overall prediction accuracy is relatively low, we can achieve a much higher prediction performance for a subgroup of users.

Given the simultaneously predicted ($t_i$,$o_i$,$d_i$), it is useful to evaluate the chances that the prediction is partially correct regarding each of the three attributes separately. Table 5 also shows the partial prediction accuracy for the three trip attributes. Note that the prediction accuracy of $o_i$ depends on $t_i$, and $d_i$ depends on $t_i$ and $o_i$. Comparing Table 5 against Table 4, it is found that simultaneous prediction achieves similar results for prediction of $t_i$, but worse results for prediction of $o_i$ and $d_i$. However, the margin is much smaller for P2B; the median prediction accuracy decreases by 2.8% for $o_i$ and 4.8% for $d_i$. Note that the difference in the prediction accuracy of $d_i$ can be eliminated if we allow prediction updates upon observing a transaction at the origin station. For intrinsic mobility data, users often generate two transaction records for every trip, one at the start of the trip and the other at the end. Instead of predicting $d_i$ simultaneously with $t_i$ and $o_i$, we may predict $d_i$ based on true $t_i$ and $o_i$ after the trip starts. The potential improvement in prediction performance is larger for P2A. As a trade-off, this may limit its implications for travel demand management (TDM). It is much more difficult to influence users' travel choices once the trip starts. On the other hand, dynamic destination prediction at the start of the trip can be useful for real-time applications such as dynamic travel time prediction and path recommendation.

## 5. Discussion

Whereas previous research focuses on predicting the next location of users, this paper presents a methodology to predict the next trip using transit smart card data. An individual's daily mobility is represented as a chain of trips, each defined as a combination of three attributes—trip start time *t*, origin *o*, and destination *d*. The method first predicts whether the user will make a trip within a day, and then, if so, predict the attributes of the next trip. Regularized logistic regression models are adopted for trip making prediction. Inspired by the *n*-gram models commonly used in natural language modeling, a mobility *n*-gram model is developed for trip attribute prediction. The proposed methodology is tested using pseudonymized transit smart card data from more than 10,000 users in London, U.K. over two years. Our trip making prediction models achieve a median prediction accuracy of greater than 80%. For trip attribute prediction, we find that (1) the first trip of the day is generally challenging to predict; (2) *t* is less predictable than *o* and *d*, implying higher intra-personal variability in temporal choices than spatial choices; (3) inter-personal variability is demonstrated through considerable variation in prediction accuracies across users, although positive correlations are found between trip attributes; (4) overall, the prediction accuracy for an average user is around 40% for *t*, 70–80% for *o*, 60–70% for *d*, and 20–30% for the combination of ($t$,$o$,$d$).

Because the full attributes of the next trip of an individual constitute a multi-dimensional distribution, the prediction of the combination of (t, o, d) is challenging. Nevertheless, certain users and certain aspects of travel behavior are more predictable than others. In practice, policies and strategies can be designed targeting at the predictable users and behavior aspects. Besides, our method provides a probability distribution over all possible outcomes of a trip. For some applications, we may need to account for a range of possibilities, based on prediction ranks (see Fig. 6), or the entire probability distribution. For example, a traveler information system may be developed to inform users about service shortage or delays in certain areas during certain time periods. This will allow users to make informed travel choices that potentially help rebalance the mismatch between travel demand and supply. Individual mobility prediction can be used to dynamically match individual users with relevant information based on the estimated probability that the user will travel to the affected areas during the affected time periods.

Even though we use the passenger railway system as the example in this paper, the method is agnostic to particular modes. The only information required is the starting time, origin, destination, and a user identifier of each trip. New mobility service providers, such as ride sharing, e-hailing services, bike sharing programs, car-sharing services as well as on-demand “pop-up” bus services, also collect individual-level travel records similar to the transit smart card data. An important feature of our method is that the data sequence does not need to be complete. The algorithm can work with any *consistent* subset of the travel sequences of a person, such as all the bicycle trips of a person; all the UBER trips of a person; and all the subway trips of a person, etc.

The proposed methodology represents a promising direction for future research, and it can be improved and expanded in several ways. Users may be clustered based on their individual mobility patterns, so that for each user we can identify other similar users whose travel records are most relevant. This helps to compensate the sparsity of individual data, which is particularly important for infrequent or new users, for whom little or no personal travel history is available. Furthermore, in the present methodology, a special smoothing strategy is applied to account for the correlation between adjacent time bands. Similar strategies could be applied to locations as well. Location adjacency may be measured using geographical distance (for predicting $o$) or network distance (for predicting $d$). Besides, based on the current results, model improvements are needed to better predict the first trip of the day. This may be done by considering cross-day effects, i.e., the impact of the last trip of a day on the first trip of the next day. Finally, other data sources (e.g., weather, events) may be incorporated in the model to improve the prediction performance. Particularly, intrinsic and extrinsic mobility data are complementary to each other; the integration of the two will likely overcome their limitations and provide a holistic picture of individual mobility.

## Acknowledgments

The authors would like to thank Transport for London for providing the financial and data support that made this research possible.

## Appendix A. Time-smoothed counting function

In the proposed model, we adopt a discrete representation of the time variable $t$. Specifically, $t$ is aggregated to hourly bands, although the choice of the bandwidth can be adjusted. A discrete representation of time is computationally convenient, but less realistic. For example, two time bands that are next to each other would be regarded as independent, despite our intuition that they should be similar. Here, we assume two contexts are adjacent if they are only one time unit apart. Formally, $x'_{1:n-1}$ is *adjacent* to $x_{1:n-1}$ if there exists some integer $z \in [1,n-1]$ such that $x'_z$ is an index of a time band and replacing $x'_z$ with $x'_z+1$ or $x'_z-1$ results in a change of contexts from $x'_{1:n-1}$ to $x_{1:n-1}$.

Let $\boldsymbol{A}_{\boldsymbol{x_{1:n-1}}}$ be the set of all contexts adjacent to $x_{1:n-1}$. Based on this definition, the time-smoothed counting function $C(x_{1:n-1},x_n)$ returns the average of the frequencies of $(x_{1:n-1},x_n)$ and $(x'_{1:n-1},x_n)$, $\forall\ x'_{1:n-1} \in \boldsymbol{A}_{\boldsymbol{x_{1:n-1}}}$, shown in Eq. (A.1). In this way, we build in correlations between adjacent time bands and smooth the frequency counts.

$$C(x_{1:n-1},x_n) = \frac{c(x_{1:n-1},x_n) + \sum_{x'_{1:n-1}\in \boldsymbol{A}_{\boldsymbol{x_{1:n-1}}}} c(x'_{1:n-1},x_n)}{1+|\boldsymbol{A}_{\boldsymbol{x_{1:n-1}}}|} \tag{A.1}$$

where $c(x_{1:n-1},x_n)$ denotes a counting function that returns the number of times $(x_{1:n-1},x_n)$ appears in the training data.

## Appendix B. Mutual information analysis of consecutive trips

Unlike the correlation coefficient that is limited to real-valued random variables, Mutual Information (MI) is more general and determines how similar the joint distribution of two random variables is to the products of their factored marginal distribution. Formally, the mutual information of two discrete random variables $X$ and $Y$ can be defined as

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2\frac{P(x,y)}{P(x)P(y)} \tag{B.1}$$

Eq. (B.1) is applied to analyze the interdependencies of the trip attributes of consecutive trips in a day, and the results are shown in Fig. B.8. Note that the mutual information of a random variable and itself is its entropy. Between the two consecutive trips, the highest mutual information is found between $o_i$ and $d_{i-1}$. Therefore, in Fig. 2 $d_{i-1}$ is placed to the right of all other context variables for prediction of $o_i$, and it indeed improves the prediction performance. The ordering of context variables in Fig. 2 is largely, albeit not completely, based on the ranking of mutual information.

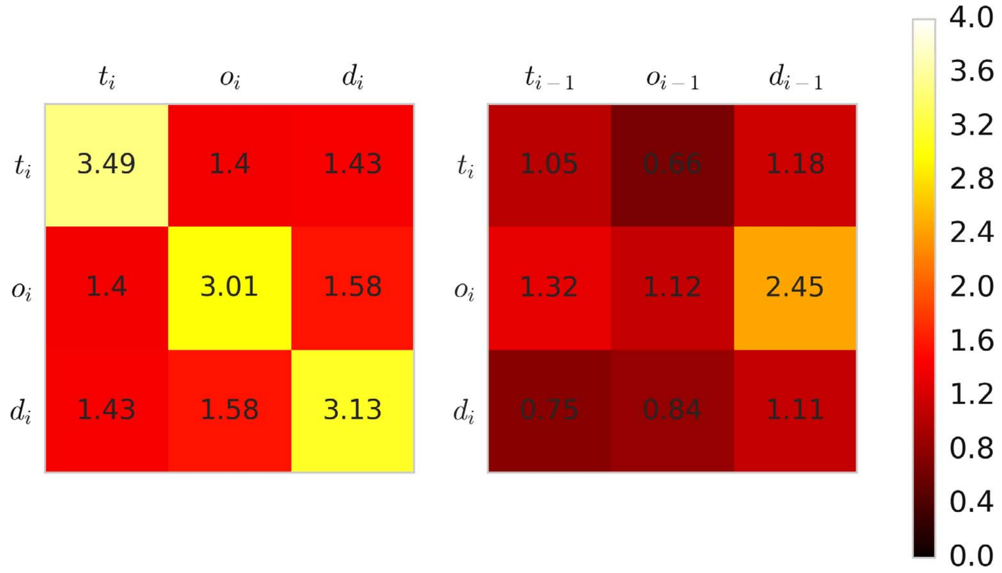

**Fig. B.8.** Mutual information of two consecutive trips.

## Appendix C. Time prediction errors

In implementation, $t$ is aggregated to hourly bands, and our prediction is considered correct only if the predicted hour band matches the true band. When a prediction is incorrect, it is useful to examine the magnitude of the error, measured as the difference between the predicted time and the true time, or $e = t_{predicted} - t_{true}$. As $t$ is measured in hours, the units of $e$ are also hours.

Fig. C.9 shows the distribution of the time prediction error. The error is narrowly centered at 0, meaning that even when the model makes an error, it is likely to be small. The mean error $E[e]$ and the mean absolute error $E[|e|]$ for P2A and P2B are given at the top right of Fig. C.9. The models make bigger errors in predicting $t_1$. Furthermore, while the mean error for P2B is close to 0, the mean error for P2A is negative. This means that the true $t_1$ is, more often than not, later than the predicted $t_1$. Note that the predicted $t_1$ is often the most frequent hour when the individual starts the first trip of the day, and for commuters the first trip of the day is usually a commuting trip. The most common hour of the day when people leave home for work tends to be earlier than the time they leave home for other purposes. Thus the actual $t_1$ could be much later than the predicted $t_1$ when a commuter takes a day off for reasons unknown to the model.

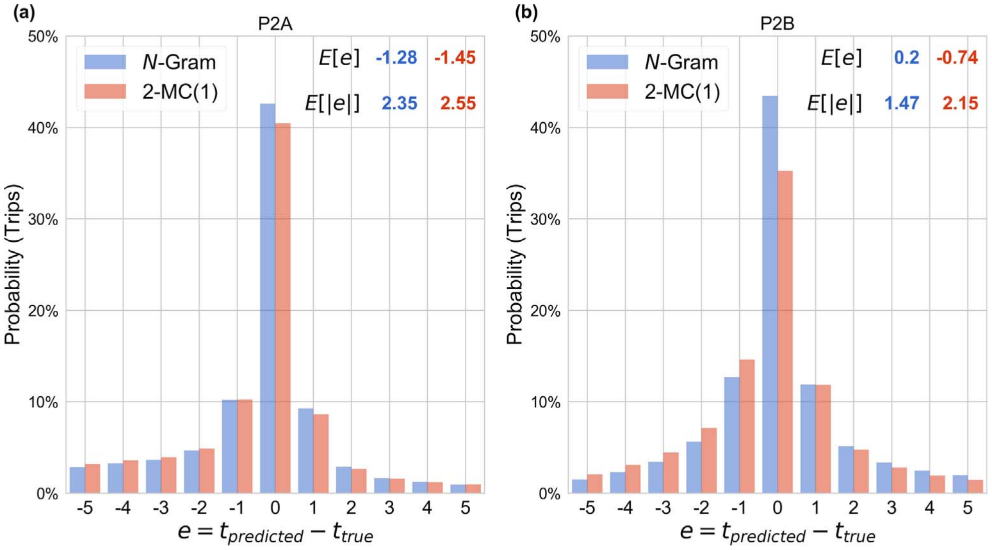

**Fig. C.9.** Distribution of time prediction error.

## Appendix D. Correlation in predictability of trip attributes

As shown in Fig. 5, the prediction accuracy for $t_i$,$o_i$, and $d_i$ varies significantly across users. Such variability may be partly explained by individual heterogeneity in travel patterns and transit use patterns. Fig. D.10 shows the correlation in prediction accuracy between different trip attributes. Overall, the correlation coefficient ranges from 0.36 to 0.56, and the correlation is stronger for P2A. This implies that users with higher predictability in one trip attribute tend to show higher predictability in other trip attributes. This is not surprising, because different travel choices of the same individual are expected to be correlated. A regular commuter who adheres to the same travel routine every day is likely to exhibit high predictability in all trip attributes.

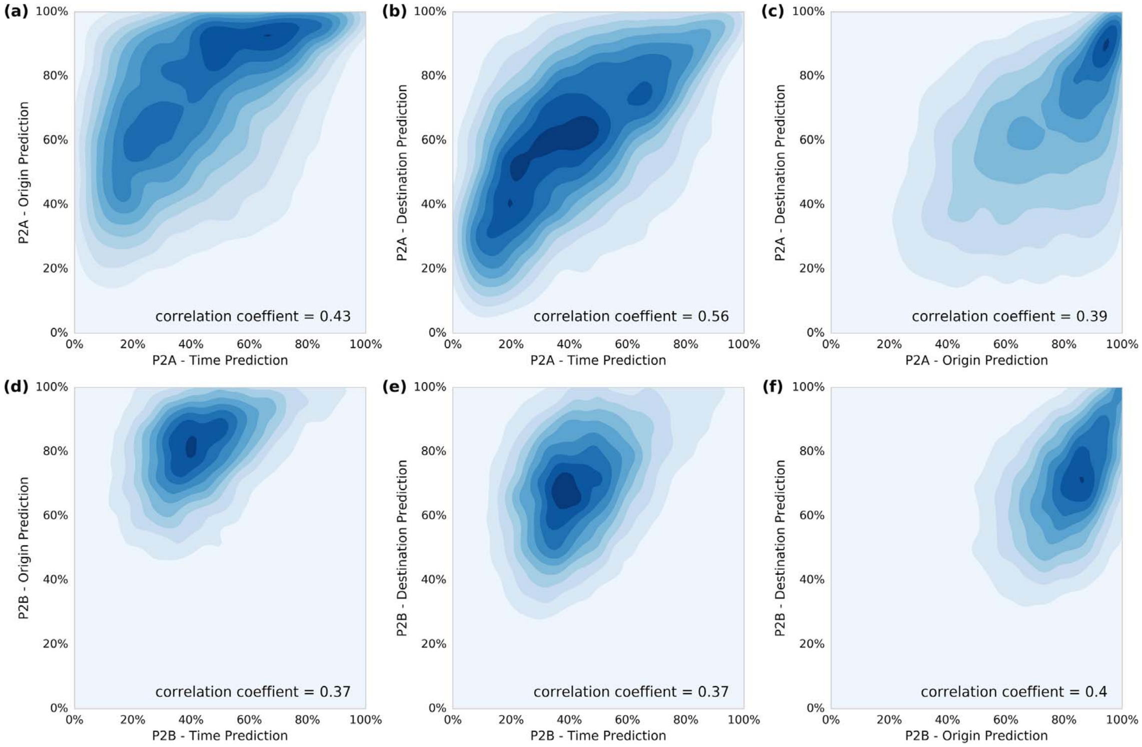

**Fig. D.10.** Correlation in prediction accuracy between different trip attributes.

## References

Alhasoun, F., Alhazzani, M., Aleissa, F., Alnasser, F., González, M., 2017. City scale next place prediction from sparse data through similar strangers. In: Proceedings of ACM KDD Workshop, Halifax, Canada, August 14, 2017 (UrbComp'17).

Asahara, A., Maruyama, K., Sato, A., Seto, K. 2011. Pedestrian-movement prediction based on mixed Markov-chain model. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems GIS '11. ACM, New York, NY, USA, pp. 25–33. http://dx.doi.org/10.1145/2093973.2093979.

Calabrese, F., Lorenzo, G.D., Ratti, C., 2010. Human mobility prediction based on individual and collective geographical preferences. In: 2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 312–317. http://dx.doi.org/10.1109/ITSC.2010.5625119.

Chen, S.F., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Comput. Speech Langu. 13, 359–394. http://dx.doi.org/10.1006/csla.1999.0128. URL <http://www.sciencedirect.com/science/article/pii/S0885230899901286> .

Colombo, G.B., Chorley, M.J., Williams, M.J., Allen, S.M., Whitaker, R.M., 2012. You are where you eat: foursquare checkins as indicators of human mobility and behaviour. In: 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 217–222. http://dx.doi.org/10.1109/PerComW.2012.6197483.

Dou, M., He, T., Yin, H., Zhou, X., Chen, Z., Luo, B., 2015. Predicting passengers in public transportation using smart card data. In: Sharaf, M.A., Cheema, M.A., Qi, J. (Eds.), Databases Theory and Applications number 9093 in Lecture Notes in Computer Science. Springer International Publishing, pp. 28–40. http://dx.doi.org/10.1007/978-3-319-19548-3_3.

Eagle, N., Pentland, A.S., 2006. Reality mining: sensing complex social systems. Pers. Ubiquit. Comput. 10, 255–268. http://dx.doi.org/10.1007/s00779-005-0046-3. URL <https://link.springer.com/article/10.1007/s00779-005-0046-3> .

Eagle, N., Pentland, A.S., 2009. Eigenbehaviors: identifying structure in routine. Behav. Ecol. Sociobiol. 63, 1057–1066. http://dx.doi.org/10.1007/s00265-009-0739-0. URL <http://link.springer.com/article/10.1007/s00265-009-0739-0> .

Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., Bento, C., 2014. Catch me if you can: predicting mobility patterns of public transport users. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1995–2002. http://dx.doi.org/10.1109/ITSC.2014.6957997.

Gambs, S., Killijian, M.-O., del Prado Cortez, M.N., 2012. Next place prediction using mobility Markov chains. In: Proceedings of the First Workshop on Measurement, Privacy, and Mobility MPM '12. ACM, New York, NY, USA, pp. 3:1–3:6. http://dx.doi.org/10.1145/2181196.2181199.

Gidófalvi, G., Dong, F., 2012. When and where next: individual mobility prediction. In: Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems MobiGIS '12. ACM, New York, NY, USA, pp. 57–64. http://dx.doi.org/10.1145/2442810.2442821.

González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782. http://dx.doi.org/10.1038/nature06958. URL <http://www.nature.com/doifinder/10.1038/nature06958> .

Goulet-Langlois, G., Koutsopoulos, H.N., Zhao, Z., Zhao, J., 2017. Measuring regularity of individual travel patterns. IEEE Trans. Intell. Transp. Syst. 1–10. http://dx.doi.org/10.1109/TITS.2017.2728704.

Hasan, S., Schneider, C.M., Ukkusuri, S.V., Gonzlez, M.C., 2012. Spatiotemporal patterns of urban human mobility. J. Stat. Phys. 151, 304–318. http://dx.doi.org/10.1007/s10955-012-0645-0. URL <http://link.springer.com/article/10.1007/s10955-012-0645-0> .

Hasan, S., Zhan, X., Ukkusuri, S.V., 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing UrbComp '13. ACM, New York, NY, USA, pp. 6:1–6:8. http://dx.doi.org/10.1145/2505821.2505823.

Hawelka, B., Sitko, I., Kazakopoulos, P., Beinat, E., 2017. Collective prediction of individual mobility traces for users with short data history. PLOS ONE 12, e0170907. http://dx.doi.org/10.1371/journal.pone.0170907. URL <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0170907>.

Hsieh, H.-P., Li, C.-T., Gao, X., 2015. T-gram: a time-aware language model to predict human mobility. In: Ninth International AAAI Conference on Web and Social Media. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10559>.

Jurafsky, D., Martin, J.H., 2008. N-grams. In: Speech and Language Processing, second ed. Prentice Hall, Upper Saddle River, N.J, pp. 83–122.

Katz, S., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. Acoust. Speech Signal Process. 35, 400–401. http://dx.doi.org/10.1109/TASSP.1987.1165125.

Kim, M., Kotz, D., 2007. Periodic properties of user mobility and access-point popularity. Pers. Ubiquit. Comput. 11, 465–479. http://dx.doi.org/10.1007/s00779-006-0093-4. URL <https://link.springer.com/article/10.1007/s00779-006-0093-4> .

Lathia, N., Smith, C., Froehlich, J., Capra, L., 2013. Individuals among commuters: building personalised transport information services from fare collection systems. Pervasive Mob. Comput. 9, 643–664. http://dx.doi.org/10.1016/j.pmcj.2012.10.007. URL <http://www.sciencedirect.com/science/article/pii/S1574119212001356> .

Lu, X., Wetter, E., Bharti, N., Tatem, A.J., Bengtsson, L., 2013. Approaching the limit of predictability in human mobility. Sci. Rep. 3. http://dx.doi.org/10.1038/srep02923. http://www.nature.com/articles/srep02923.

MacKay, D.J.C., Peto, L.C.B., 1994. A hierarchical dirichlet language model. Nat. Lang. Eng. 1, 1–19.

Mathew, W., Raposo, R., Martins, B., 2012. Predicting future locations with hidden Markov models. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing UbiComp '12. ACM, New York, NY, USA, pp. 911–918. http://dx.doi.org/10.1145/2370216.2370421. URL http://doi.acm.org/10.1145/2370216.2370421.

Ng, A.Y., 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proceedings of the Twenty-first International Conference on Machine Learning ICML '04. ACM, New York, NY, USA, pp. 78. http://dx.doi.org/10.1145/1015330.1015435. URL http://doi.acm.org/10.1145/1015330.1015435.

Noulas, A., Scellato, S., Lathia, N., Mascolo, C., 2012. Mining user mobility features for next place prediction in location-based services. In: International Conference on Data Mining. IEEE.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.-L., 2015. Returners and explorers dichotomy in human mobility. Nat. Commun. 6, 8166. http://dx.doi.org/10.1038/ncomms9166. URL <http://www.nature.com/ncomms/2015/150908/ncomms9166/full/ncomms9166.html> .

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238. http://dx.doi.org/10.1109/TPAMI.2005.159.

Purnama, I.B.I., Bergmann, N., Jurdak, R., Zhao, K., 2015. Characterising and predicting urban mobility dynamics by mining bike sharing system data. In: 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), pp. 159–167. http://dx.doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.46.

Sapiezynski, P., Stopczynski, A., Gatej, R., Lehmann, S., 2015. Tracking human mobility using WiFi signals. PLoS ONE, 10. http://dx.doi.org/10.1371/journal.pone.0130824. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4489206/.

Schneider, C.M., Belik, V., Couronne, T., Smoreda, Z., Gonzalez, M.C., 2013. Unravelling daily human mobility motifs. J. Roy. Soc. Interface 10, 20130246. http://dx.doi.org/10.1098/rsif.2013.0246. URL <http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2013.0246> .

Song, C., Koren, T., Wang, P., Barabási, A.-L., 2010a. Modeling the scaling properties of human mobility. Nat. Phys. 6, 818–823. http://dx.doi.org/10.1038/nphys1760.. Available from: < 1010.0436 > .

Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010b. Limits of predictability in human mobility. Science 327, 1018–1021. http://dx.doi.org/10.1126/science.1177170. URL <http://www.sciencemag.org/content/327/5968/1018> .

Teh, Y.W., 2006. A hierarchical bayesian language model based on pitman-yor processes. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics ACL-44. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 985–992. http://dx.doi.org/10.3115/1220175.1220299.

Zhao, K., Musolesi, M., Hui, P., Rao, W., Tarkoma, S., 2015. Explaining the power-law distribution of human mobility through transportation modality decomposition. Sci. Rep. 5, 9136. http://dx.doi.org/10.1038/srep09136. URL <http://www.nature.com/srep/2015/150316/srep09136/full/srep09136.html> .

Zhao, Z., Zhao, J., Koutsopoulos, H.N., 2016. Individual-Level Trip Detection using Sparse Call Detail Record Data based on Supervised Statistical Learning. URL <https://trid.trb.org/view.aspx?id=1393647>.

Zhong, C., Manley, E., Mller Arisona, S., Batty, M., Schmitt, G., 2015. Measuring variability of mobility patterns from multiday smart-card data. J. Comput. Sci. 9, 125–130. http://dx.doi.org/10.1016/j.jocs.2015.04.021. URL <http://www.sciencedirect.com/science/article/pii/S1877750315000599> .