Discovering latent activity patterns from transit smart card data: A spatiotemporal topic model

Discovering Latent Activity Patterns from Transit Smart Card Data: A Spatiotemporal Topic Model

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*

Abstract

Although automatically collected human travel records can accurately capture the time and location of human movements, they do not directly explain the hidden semantic structures behind the data, e.g., activity types. This work proposes a probabilistic topic model, adapted from Latent Dirichlet Allocation (LDA), to discover representative and interpretable activity categorization from individual-level spatiotemporal data in an unsupervised manner. Specifically, the activity-travel episodes of an individual user are treated as words in a document, and each topic is a distribution over space and time that corresponds to certain type of activity. The model accounts for a mixture of discrete and continuous attributes—the location, start time of day, start day of week, and duration of each activity episode. The proposed methodology is demonstrated using pseudonymized transit smart card data from London, U.K. The results show that the model can successfully distinguish the three most basic types of activities—home, work, and other, and it fits the data significantly better than rule-based approaches. As the specified number of activity categories increases, more specific subpatterns for home and work emerge. This work makes it possible to enrich human mobility data with representative and interpretable activity patterns without relying on predefined activity categories or heuristic rules.

Keywords: Human mobility, Activity discovery, Spatiotemporal pattern, Topic model, Transit smart card

1. Introduction

The spatiotemporal aspect of our lives can be segmented into episodes of travel and activity participation. Activities have long been recognized as the fundamental driver of travel demand. In activity-based analysis of travel behavior, travel is treated as being derived from the need to pursue activities distributed in space (Axhausen and Gärling, 1992; Bhat and Koppelman, 1999; Bowman and Ben-Akiva, 2001; Rasouli and Timmermans, 2014). A *trip* is defined as "the travel required from an origin location to access a destination for the purpose of performing some activity" (McNally, 2007), and an *activity episode* refers to a discrete activity participation (time allocated to activities) at a location (Bhat and Koppelman, 1999). By definition, each trip is followed by an activity episode, and the attributes of the trip are determined based on the activity participation at the trip destination. Therefore, individual mobility is closely intertwined with activity participation. Understanding activity patterns has important applications in urban and transportation planning, location-based services, public health and safety, and emergency response.

Recent years have seen an explosion of large-scale spatiotemporal datasets related to human mobility, such as cellular network data, transit smart card data, and geo-tagged social media data. Although such automated data sources can capture the time and location of some human mobility with precision and at a fine level of detail, they do not explicitly provide any behavioral explanation, e.g., why people visit a certain place at a certain time. Traditionally, the most common way to collect such information is through manual surveys of individual activity participation, which are costly and do not scale well. A number of methods have been proposed to infer the activity based on heuristic rules (Alexander et al., 2015; Zou et al., 2018) , and/or supervised learning models fitted using the survey data (Liao et al., 2005; Allahviranloo and Recker, 2013). Both require predefined activity categories (e.g., home, work, school, recreation) that are often come up by the researchers. However, it is debatable whether such categorization is truly representative of the richness and diversity of human activities. Specifically, for human mobility research, we are most interested in finding the types of activities that drive distinctive spatiotemporal travel behavior. In this work, we focus on *activity discovery* (i.e., finding representative activity categories) instead of *activity inference* (i.e., predicting predefined activity categories). Of course, the two tasks are closely connected. Analyzing discovered activity patterns can help researchers design better rules to infer them.

Automatic activity discovery is a challenging task, as people's spatiotemporal choices vary from day to day and from individual to individual. Some of the variations can be explained by different underlying activities (i.e., inter-activity variability), and some are attributed to exogenous factors (e.g., weather) and thus become inherent randomness for the same activity (i.e., intra-activity variability). Longitudinal spatiotemporal data itself generally contains a significant amount of structure (Eagle and Pentland, 2009). Assuming that people's spatiotemporal choices for each activity episode are generated based on the specific activity they intend to participate in, it is possible to find the latent activity patterns that underlie human mobility. This would require an unsupervised approach that is able to sift through large amounts of noisy data and find meaningful underlying activities. Unlike supervised learning, it does not require training data, and has the potential of automatic discovery of emerging activity patterns (Farrahi and Gatica-Perez, 2009, 2011; Hasan and Ukkusuri, 2014). The objective of this study is to develop a methodology that can help us uncover the latent activity patterns from large-scale human mobility datasets.

In this work, we propose a model that extends Latent Dirichlet Allocation (LDA), a well known probabilistic topic model first introduced by Blei et al. (2003). Topic models are generative models that represent documents as mixtures of topics, and assign a topic to each word in a document. As this representation shares some similarities with individual mobility, as shown in Table 1, it can be adapted for latent activity discovery. In the proposed model, we treat the activity-travel history of each individual as a document, and each activity episode as a *multi-dimensional* word. This would allow us to discover the latent activity associated with each activity episode and the activity mixture with each individual, based on the spatiotemporal data observed. The discovered activity patterns can then be used to understand time allocation behavior, predict human mobility, and characterize urban land uses.

Table 1: Related concepts in natural language and human mobility

Natural language terminology	Human mobility terminology	General terminology
Word	Activity episode (or trip)	Observation
Document	Individual travel-activity history	Group of observations
Topic	Activity	Latent component

The paper has two main contributions:

- We demonstrate that topic models can be extended for latent activity discovery at the individual trip (or activity episode) level based on unannotated travel records. This is distinctly different from previous studies that have applied topic models for discovery of daily or weekly activity patterns based on annotated data (Farrahi and Gatica-Perez, 2009; Hasan and Ukkusuri, 2014). Without activity labels provided in the unannotated data, one can only directly use the high-dimensional spatio-temporal information, which makes the problem more challenging.
- The proposed methodology presents a flexible way to combine continuous time variables and discrete location variables for latent activity discovery. In contrast, existing methods mostly rely on the discretized representation of time (Hasan and Ukkusuri, 2014; Sun and Axhausen, 2016; Sun et al., 2019). The continuous representation of time not only better reflects people's actual temporal preferences, but also mitigates data sparisity. In particular, we show that the use of activity duration, along with start time and location of the activity episode, greatly enhance the interpretability of the discovered latent activity patterns.

2. Literature Review

A plethora of methods have been proposed in the literature for activity inference. They can be generally categorized into two types—rule-based methods, and model-based methods. In rule-based methods, heuristic decision rules and thresholds are specified by researchers to categorically determine the activity. For example, based on Alexander et al. (2015), an individual's home location is identified as the stay with the most visits on weekends and weekdays between 7 pm and 8 am. Hasan et al. (2013) assumed that one's home and workplace were the most and second most visited places, respectively. Also based on transit smart card data, Zou et al. (2018) proposed a more complicated decision process that considered the time, location, card type, and travel regularity. While these rule-based methods have been shown to work well in practice, they require domain knowledge to design the rules and do not provide an estimation of uncertainty. More importantly, one implicit assumption of most rule-based methods is that the activity is uniquely determined based on the location, i.e., there can only be one activity performed in a location. This is probably not true, especially for dense urban areas with highly mixed land use.

Model-based activity inference overcomes many limitations of rule-based methods, but the true activities associated with travel records need to be provided. For example, using annotated GPS data, Liao et al. (2005) proposed a new approach for activity inference based on Relational Markov Networks (RMN) and Conditional Random Fields (CRF). Allahviranloo and Recker (2013) adopted a multi-class Support Vector Machine (SVM) approach to infer the activity type, and validated it on a subset of the 2001 California Personal Travel Survey data. More recently, researchers turned to data fusion to form labeled training samples. This was commonly done by combining mobility data (e.g., transit smart card data) with survey data (Lee and Hickman, 2014; Kusakabe and Asakura, 2014; Alsger et al., 2018). The advancement of information and communication technologies has made data fusion more feasible. For example, Kim et al. (2014) demonstrated the feasibility of activity inference using data from the Future Mobility Survey (FMS), a smartphone based activity-travel survey system, which acquires movement data through sensors in smartphones and activity information through a web-based interactive process. Despite of the improved model performance, these methods still depend on predefined activity categorization. A more fundamental problem is how to find the right activity categorization.

For activity discovery, the activity information is not provided, and the problem is to discover and interpret latent patterns from the data. In one of the first studies of this kind, Eagle and Pentland (2009) used Principle Component Analysis (PCA) to extract a set of characteristic behavior vectors, called "eigenbehavior" from mobile phone data. Apart from PCA, other variations of dimension reduction methods have been applied to discover latent patterns from human mobility data, including non-negative matrix factorization (Peng et al., 2012), and probabilistic tensor factorization (Sun and Axhausen, 2016). A Continuous Hidden Markov Model (CHMM) was proposed in Han and Sohn (2016) to impute the sequence of activities for each trip chain. Overall, these methods are not suitable for grouped data, where multiple trips associated with the same individual are highly correlated. As activity patterns vary across individuals, it is important to account for heterogenous behavior at the individual level. To address this issue, a hierarchical structure may be adopted, which would capture both inter-individual and intra-individual variations at different levels in the hierarchy.

First introduced by Blei et al. (2003), Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of grouped discrete data. Each group is described as a random mixture over a set of latent topics where each topic is a discrete distribution over the collection's vocabulary. Other more recent topic models are generally extensions of LDA, including the dynamic topic model (Blei and Lafferty, 2006), supervised topic model Blei and McAuliffe (2010), and Hierarchical Dirichlet Process (HDP) (Teh et al., 2006). Originally designed as a text mining tool, it has found application in other fields such as image processing (Rasiwasia and Vasconcelos, 2013) and bioinformatics (Liu et al., 2016). In transportation research, it has been used for mining transportation-related social media posts (Hidayatullah and Ma'arif, 2017), and understanding driving states (Chen et al., 2019), and extracting spatiotemporal patterns in bikesharing systems (Côme et al., 2014; Montoliu, 2012). Sun et al. (2019) adapted LDA for spatiotemporal data and tested it on license plate recognition data. For activity discovery, it was first applied to wearable sensor data in Huynh et al. (2008). Regarding its application to mobility analysis, Farrahi and Gatica-Perez (2009, 2011) adapted the LDA model for annotated mobile phone data, in which the daily mobility of an individual is represented as a "bag of location sequences". Later, a similar approach was used by Hasan and Ukkusuri (2014) to find weekly activity patterns from individual activity information shared in social media. All of these studies focus on identifying routines (or combinations of activities over a time period) based on annotated activity data. Under this problem definition, each topic represents a distinct distribution over activity sequences (Farrahi and Gatica-Perez, 2009) or timestamped activities (Hasan and Ukkusuri, 2014). In contrast, our work focuses on identifying activities from travel records, where each topic is a distinct distribution over time and space. There is a significant difference in problem dimensionality; there are typically many more locations than activity categories. The need to work with high-dimensional location data, in combination with sparsity of the data (compared to text data), makes it difficult to directly apply traditional LDA model for our problem.

Another major difference lies in how we represent time. Most prior studies (Hasan and Ukkusuri, 2014; Sun and Axhausen, 2016; Sun et al., 2019) used discretized representation of time. This is obviously not ideal, as the boundaries we choose to divide time are usually arbitrary and do not perfectly capture people's temporal preferences. In addition, discretized representation of time makes it more challenging to discover meaningful patterns with limited data, especially when the number of time categories is high, e.g., one category for each hour of the week (Hasan and Ukkusuri, 2014). To address these issues, we choose to represent time with three different variables—day of the week, time of day, and duration, of which the latter two are continuous. This not only offers a more natural representation of people's temporal behavior, and but also mitigates the data sparsity problem. The next section will present an extended LDA model that makes it possible to combine multi-dimensional and heterogeneous spatiotemporal data, for the purpose of discovering latent activity patterns.

A similar approach was proposed by Zheng et al. (2014) for mobile context discovery. It considered both spatial and temporal aspects of human behavior, but focused on identifying temporal routines. Specifically, the spatial patterns were forced to be individual-specific and could not be shared across individuals. This may limit the method's ability to uncover activities based on land use patterns. The method was validated with detailed mobile phone data from 20 participants with complete survey information. For large-scale application, however, such detailed information is rarely available. Despite of the similarity, this work can be distinguished in several ways. First, both spatial and temporal patterns are treated as global; they can be shared across individuals. In this work, each "topic" is a latent activity characterized by a distinct spatiotemporal distribution. Second, the duration of an activity episode is included in this analysis, which provides valuable information for activity discovery and interpretation. Third, for the arrival time and the duration of an activity episode, their variances are allowed to vary across activities, representing different temporal flexibilities. For example, work activities typically are less flexible than recreational activities. Fourth, the proposed methodology is validated using a large collection of individual-level transit smart card records. Unlike mobile phone data, transit smart card data is intrinsic to human mobility (Zhao et al., 2018b). As a result, the model needs to be adapted to match the characteristics of the data.

3. Methodology

3.1. Problem Formulation

Let us assume that for each individual m ($m = 1, ..., M$), we observe a collection of N_m trips, each followed by an activity episode, and the n-th trip (or activity episode) of individual m is associated with a latent activity z_{mn}. Only the spatiotemporal attributes of the activity episodes are observable. The goal is to find z_{mn} that can best explain the data.

To reflect individual heterogeneity, z_{mn} is assumed to follow an individual-specific categorical distribution parameterized by π_m. In other words, different individuals may have different composition of activities. For example, some individuals travel mainly for commuting, while others for recreation. π_m may be used to characterize the activity patterns of individual m.

Each activity episode is characterized by a set of spatiotemporal attributes, which should be chosen based on the problem and the available data source. For the purpose of latent activity discovery, we should choose the attributes that can help distinguish between different activities. In this study, we consider four attributes: the location x_{mn}, arrival time t_{mn}, day of week d_{mn}, and duration r_{mn} (i.e., how long the activity episode lasts). Both d_{mn} and x_{mn} are discrete, but t_{mn} and r_{mn} are continuous variables. Based on the activity-based analysis framework, the distributions of these variables depend on z_{mn}. For this problem, x_{mn} and d_{mn} conditional on z_{mn} are assumed to follow a categorical distribution parameterized by θ_z and ϕ_z respectively. t_{mn} is assumed to follow a normal distribution parameterized by mean μ_z and precision τ_z. Unlike arrival time, the distribution of duration is bounded on the left (i.e., nonnegative) and heavy-tailed on the right. Therefore, r_{mn} is assumed to follow a log-normal distribution parameterized by η_z and λ_z.

Bayesian inference and conjugate priors are commonly used for estimating distribution parameters from data. Based on Bayesian inference, we can update our knowledge of a parameter by incorporating new observations. The use of conjugate priors allows all the results to be derived in closed form. In this study, the prior distribution of π_m, θ_z, and ϕ_z is assumed to be a Dirichlet, which is the conjugate prior distribution of the categorical distribution. Both (μ_z, τ_z) and (η_z, λ_z) are assumed to be sampled from a normal-gamma distribution, which is the conjugate prior of the normal distribution with unknown mean and precision. These prior distributions have hyperparameters that need to be chosen by researchers.

Specifically, the proposed model assumes the data are generated according to the following process:

1. For each activity $z = 1, 2, ..., Z$,
   (a) Sample a location distribution $\theta_z \sim \text{Dirichlet}(\beta)$
   (b) Sample a day of week distribution $\phi_z \sim \text{Dirichlet}(\gamma)$
   (c) Sample a time of day distribution $\mu_z, \tau_z \sim \text{NormalGamma}(\mu_0, \kappa_0, \epsilon_0, \tau_0)$
   (d) Sample a duration distribution $\eta_z, \lambda_z \sim \text{NormalGamma}(\eta_0, \nu_0, \omega_0, \lambda_0)$
2. For each individual $m = 1, 2, ..., M$,
   (a) Sample an activity distribution: $\pi_m \sim \text{Dirichlet}(\alpha)$
   (b) For each activity episode of the individual $n = 1, 2, ..., N_m$,
      i. Sample an activity $z_{mn} \sim \text{Categorical}(\pi_m)$
      ii. Sample a location $x_{mn} \sim \text{Categorical}(\theta_{z_{mn}})$
      iii. Sample a day of week $d_{mn} \sim \text{Categorical}(\phi_{z_{mn}})$
      iv. Sample a time of day $t_{mn} \sim \text{Normal}(\mu_{z_{mn}}, \tau_{z_{mn}})$
      v. Sample a duration $r_{mn} \sim \text{LogNormal}(\eta_{z_{mn}}, \lambda_{z_{mn}})$

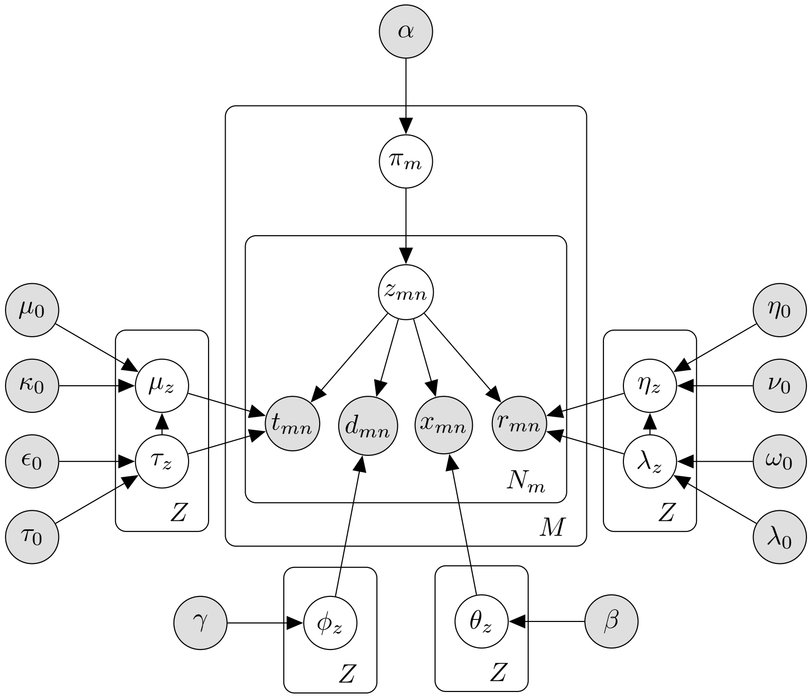


Figure 1: Plate notation of the human mobility LDA model

The structure of the adapted LDA model is shown in Figure 1, where the shaded circles represent the observed or pre-specified variables, and the non-shaded circles represent the latent variables to be estimated. The notation used in this paper is summarized in Table 2. Given hyperparameters α, β, γ, μ_0, κ_0, ϵ_0, τ_0, η_0, ν_0, ω_0, and λ_0, the generative process described above results in the following joint distribution:

$$\begin{aligned} &P(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{r}, \boldsymbol{z}, \pi, \theta, \phi, \mu, \tau, \eta, \lambda) \\ =&P(\boldsymbol{z} \mid \pi)P(\boldsymbol{x} \mid \theta_z)P(\boldsymbol{d} \mid \phi_z)P(\boldsymbol{t} \mid \mu_z, \tau_z)P(\boldsymbol{r} \mid \eta_z, \lambda_z)P(\pi)P(\theta)P(\phi)P(\mu, \tau)P(\eta, \lambda) \end{aligned} \tag{1}$$

Table 2: Notation

Notation	Explanation	Data Type
M	number of individuals	scalar
Z	number of activities	scalar
X	number of locations	scalar
D	number of days of week	scalar
N	total number of observations	scalar
N_m	number of observations for individual m	scalar
x_{mn}	location indicator for the n-th observation of individual m	scalar
d_{mn}	arrival day of week indicator for the n-th observation of individual m	scalar
t_{mn}	arrival time of day indicator for the n-th observation of individual m	scalar
r_{mn}	duration indicator for the n-th observation of individual m	scalar
z_{mn}	activity assignment indicator for the n-th observation of individual m	scalar
π_m	probabilities of z_{mn} for individual m	Z-vector
θ_z	probabilities of x_{mn} for activity z	X-vector
ϕ_z	probabilities of d_{mn} for activity z	D-vector
μ_z, τ_z	mean and precision of t_{mn} for activity z	scalar
η_z, λ_z	mean and precision of $\log(r_{mn})$ for activity z	scalar
α	Dirichlet hyperparameter for π_m	Z-vector
β	Dirichlet hyperparameter for θ_z	X-vector
γ	Dirichlet hyperparameter for ϕ_z	D-vector
μ_0, κ_0, ϵ_0, τ_0	normal-gamma hyperparameters for μ_z and τ_z	scalar
η_0, ν_0, ω_0, λ_0	normal-gamma hyperparameters for η_z and λ_z	scalar
n_z	number of observations assigned to activity z	scalar
u_{mz}	number of observations with individual m and activity z	scalar
v_{zx}	number of observations with location x and activity z	scalar
w_{zd}	number of observations with day of week d and activity z	scalar
s_z	sum of t for observations assigned to activity z	scalar
S_z	sum of t^2 for observations assigned to activity z	scalar
q_z	sum of $\log(r)$ for observations assigned to activity z	scalar
Q_z	sum of $\log(r)^2$ for observations assigned to activity z	scalar

where $\boldsymbol{x}$, $\boldsymbol{d}$, $\boldsymbol{t}$, and $\boldsymbol{r}$ are observed, and $\boldsymbol{z}$, π, θ, ϕ, μ, τ, η, and λ are latent variables to be estimated. The hyperparameters are omitted for clarity.

It is worth noting that the proposed model makes two simplifying assumptions about the structure of activity episodes. First, the sequential dependency between consecutive activity episodes are ignored. To account for the sequential dependency, we need to estimate the transition probabilities between activities, which will be difficult when the number of activities is large. In addition, it requires that the data capture a complete sequence of activity episodes, i.e., no missing activity episode is allowed, which limits the applicability of the model. In text mining, the LDA model has been proven to work well even without considering the sequential dependency across words in documents (known as "bag-of-words" assumption). Second, the distributions of different spatiotemporal attributes are assumed to be independent conditional on the activity. Estimating a joint distribution of multiple continuous and discrete variables is known to be a challenging problem. The conditional independence assumption allows us to avoid this problem and instead estimate multiple marginal distributions separately. Overall, these assumptions, although not very realistic, reduce the complexity of the model so that the latent parameters can be learned given a reasonable amount of data.

3.2. Likelihoods

To evaluate the goodness of fit of the model $\mathcal{M}$, we use the likelihood function, which can be expressed as

$$\mathcal{L}(\mathcal{M}) = P(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{r} \mid \mathcal{M}) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \sum_{z_{mn}=1}^{Z} P(z_{mn}, x_{mn}, d_{mn}, t_{mn}, r_{mn}) \tag{2}$$

For the n-th activity episode of the m-th individual, the joint probability $P(z_{mn} = z, x_{mn} = x, d_{mn} = d, t_{mn} = t, r_{mn} = r)$ can be further expanded as

$$\begin{aligned}
&\int_{\pi_m}\int_{\theta}\int_{\phi}\int_{\mu}\int_{\eta} P(\pi_m)P(\theta)P(\phi)P(\mu)P(\eta,\lambda)P(z,x,d,t,r \mid \pi_m,\theta,\phi,\mu,\eta,\lambda)\\
=&\Big(\int_{\pi_m} P(z \mid \pi_m)P(\pi_m)\Big)\cdot\Big(\int_{\theta} P(x \mid \theta_z)P(\theta)\Big)\cdot\Big(\int_{\phi} P(d \mid \phi_z)P(\phi)\Big)\\
&\cdot\Big(\int_{\mu,\tau} P(t \mid \mu_z,\tau_z)P(\mu,\tau)\Big)\cdot\Big(\int_{\eta,\lambda} P(r \mid \eta_z,\lambda_z)P(\eta,\lambda)\Big)\\
=&\frac{u_{mz}+\alpha_z}{\sum_{k=1}^{Z} u_{mk}+\alpha_k}\cdot\frac{v_{zx}+\beta_x}{\sum_{k=1}^{X} v_{zk}+\beta_k}\cdot\frac{w_{zd}+\gamma_d}{\sum_{k=1}^{D} w_{zk}+\gamma_k}\\
&\cdot\mathcal{T}\left(t \mid 2\epsilon_0+n_z, \frac{s_z+\kappa_0\mu_0}{n_z+\kappa_0}, \frac{(\tau_0+\frac{n_zS_z-s_z^2}{2n_z}+\frac{\kappa_0(s_z-n_z\mu_0)^2}{2n_z(\kappa_0+n_z)})(\kappa_0+n_z)}{(\epsilon_0+n_z/2)(\kappa_0+n_z)}\right)\\
&\cdot\mathcal{T}\left(\log(r) \mid 2\omega_0+n_z, \frac{q_z+\nu_0\eta_0}{n_z+\nu_0}, \frac{(\lambda_0+\frac{n_zQ_z-q_z^2}{2n_z}+\frac{\nu_0(q_z-n_z\eta_0)^2}{2n_z(\nu_0+n_z)})(\nu_0+n_z)}{(\omega_0+n_z/2)(\nu_0+n_z)}\right)
\end{aligned} \tag{3}$$

where the first term represents the likelihood of activity assignments, and the second through fifth terms indicate the marginal likelihood of location, day of week, time of day, and duration of stay choices given activity assignments. $\mathcal{T}(e \mid \nu, \mu, \sigma^2)$ represents the probability density function (pdf) for a generalized t-distribution with ν degrees of freedom, location parameter μ, and scale parameter σ^2. The pdf can be expressed as:

$$\mathcal{T}(e \mid \nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}}\left(1 + \frac{(e-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \tag{4}$$

Perplexity is a standard metric in machine learning to measure the performance of a probabilistic model, and it has often been used to evaluate topic models such as LDA (Farrahi and Gatica-Perez, 2011; Hasan and Ukkusuri, 2014). A lower perplexity value indicates better model performance. Perplexity can be directly calculated based on the likelihood function:

$$Perplexity = \exp\left(-\frac{\log(\mathcal{L}(\mathcal{M}))}{N}\right) \tag{5}$$

where N is the total number of activity episodes in the data.

3.3. Inference via Gibbs Sampling

In the literature, two types of approximate techniques have been adopted to estimate the LDA model—variational inference (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2004). The latter is used in this work, because it is more flexible and easier to implement. Gibbs sampling is a special case of the Markov Chain Monte Carlo (MCMC) methods, which can emulate the target posterior distribution by the stationary behavior of a Markov chain. In high-dimension cases, Gibbs sampling works by sampling each dimension iteratively, conditioned on the values of all other dimensions.

In practice, only $\boldsymbol{x}$, $\boldsymbol{d}$, $\boldsymbol{t}$, and $\boldsymbol{r}$ are observed, and we want to estimate latent variables $\boldsymbol{z}$, π, θ, ϕ, μ, τ, η, and λ. However, the latter seven variables may be integrated out, because they can be derived using the activity variable $\boldsymbol{z}$:

$$\pi_{mz} = \frac{u_{mz} + \alpha_z}{\sum_{k=1}^{Z} u_{mk} + \alpha_k} \tag{6}$$

$$\theta_{zx} = \frac{v_{zx} + \beta_x}{\sum_{k=1}^{X} v_{zk} + \beta_k} \tag{7}$$

$$\phi_{zd} = \frac{w_{zd} + \gamma_d}{\sum_{k=1}^{D} w_{zk} + \gamma_k} \tag{8}$$

$$\tau_z = \frac{\epsilon_0 + \frac{n_z}{2}}{\tau_0 + \frac{n_z S_z - s_z^2}{2n_z} + \frac{\kappa_0(s_z - n_z\mu_0)^2}{2n_z(\kappa_0 + n_z)}} \tag{9}$$

$$\mu_z = \frac{\kappa_0 + s_z}{\kappa_0 + n_z} \tag{10}$$

$$\lambda_z = \frac{\omega_0 + \frac{n_z}{2}}{\lambda_0 + \frac{n_z Q_z - q_z^2}{2n_z} + \frac{\nu_0(q_z - n_z\eta_0)^2}{2n_z(\nu_0 + n_z)}} \tag{11}$$

$$\eta_z = \frac{\nu_0 + q_z}{\nu_0 + n_z} \tag{12}$$

The strategy of integrating out some of the parameters for model inference is often referred to as *collapsed* Gibbs sampling. In order to construct a collapsed Gibbs sampler, we need to compute the probability of an activity being assigned to an observation, given all other activity assignments to all other observations. This requires the derivation of the full conditional activity distribution for a specific activity episode. Assuming that $x_{mn} = x$, $d_{mn} = d$, $t_{mn} = t$, and $r_{mn} = r$, the conditional probability of $z_{mn} = z$ is given by

$$
\begin{aligned}
&P(z_{mn} = z \mid \boldsymbol{z}^{-mn}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{r}) \\
\propto &P(z_{mn} = z, x_{mn} = x, d_{mn} = d, t_{mn} = t, r_{mn} = r \mid \boldsymbol{z}^{-mn}, \boldsymbol{x}^{-mn}, \boldsymbol{d}^{-mn}, \boldsymbol{t}^{-mn}) \\
\propto &P(z_{mn} = z \mid \boldsymbol{z}^{-mn}) \cdot P(x_{mn} = x \mid z_{mn} = z, \boldsymbol{z}^{-mn}, \boldsymbol{x}^{-mn}) \cdot P(d_{mn} = d \mid z_{mn} = z, \boldsymbol{z}^{-mn}, \boldsymbol{d}^{-mn}) \\
&\cdot P(t_{mn} = t \mid z_{mn} = z, \boldsymbol{z}^{-mn}, \boldsymbol{t}^{-mn}) \cdot P(r_{mn} = r \mid z_{mn} = z, \boldsymbol{z}^{-mn}, \boldsymbol{r}^{-mn}) \\
\propto &\frac{u_{mz}^{-mn} + \alpha_z}{\sum_{k=1}^{Z} u_{mk}^{-mn} + \alpha_k} \cdot \frac{v_{zx}^{-mn} + \beta_x}{\sum_{k=1}^{X} v_{zk}^{-mn} + \beta_k} \cdot \frac{w_{zd}^{-mn} + \gamma_d}{\sum_{k=1}^{D} w_{zk}^{-mn} + \gamma_k} \\
&\cdot \mathcal{T}\left(t \mid 2\epsilon_0 + n_z^{-mn}, \frac{s_z^{-mn} + \kappa_0\mu_0}{n_z^{-mn} + \kappa_0}, \frac{(\tau_0 + \frac{n_z^{-mn} S_z^{-mn} - (s_z^{-mn})^2}{2n_z^{-mn}} + \frac{\kappa_0 (s_z^{-mn} - n_z^{-mn}\mu_0)^2}{2n_z^{-mn}(\kappa_0 + n_z^{-mn})})(\kappa_0 + n_z^{-mn})}{(\epsilon_0 + n_z^{-mn}/2)(\kappa_0 + n_z^{-mn})} \right) \\
&\cdot \mathcal{T}\left(\log(r) \mid 2\omega_0 + n_z^{-mn}, \frac{q_z^{-mn} + \nu_0\eta_0}{n_z^{-mn} + \nu_0}, \frac{(\lambda_0 + \frac{n_z^{-mn} Q_z^{-mn} - (q_z^{-mn})^2}{2n_z^{-mn}} + \frac{\nu_0 (q_z^{-mn} - n_z^{-mn}\eta_0)^2}{2n_z^{-mn}(\nu_0 + n_z^{-mn})})(\nu_0 + n_z^{-mn})}{(\omega_0 + n_z^{-mn}/2)(\nu_0 + n_z^{-mn})} \right)
\end{aligned}
\tag{13}
$$

where the superscript $^{-mn}$ signifies leaving the n-th observation of the m-th individual out of the calculation. Note that Eq. (13) is similar to Eq. (3), which is not surprising. The probability of an activity assignment is proportional to the joint probability of the data with the activity assignment.

In practice, it is more convenient to store the input data $\boldsymbol{x}$, $\boldsymbol{d}$, $\boldsymbol{t}$, $\boldsymbol{r}$ in arrays, so that x_i, d_i, t_i, and r_i are the attributes of the i-th observation in the dataset. In order to keep track of the individual that each observation belongs to, we use another array $\boldsymbol{m}$, where m_i indicates the individual ID associated with the i-th observation. See Algorithm 1 for the detailed Gibbs Sampling procedure.

3.4. Hyperparameters

The choice of hyperparameters can significantly influence the behavior of the model. This section gives an overview of the meaning of the hyperparameters and the specific choices for this analysis.

3.4.1. Dirichlet Priors

Typically, symmetric Dirichlet priors are used in LDA, which means that the a priori assumption is that all possible outcomes have the same chance of occurring. The Dirichlet hyperparameters generally have a smoothing effect on multinomial parameters. Lowering the values of these hyperparameters will reduce the smoothing effect and increase sparsity of the posterior distribution. In the proposed model, the sparsity of the π_m, θ_z, and ϕ_z are controlled by α, β, and γ, respectively. A sparser π_m means that the model prefers to

Algorithm 1: Adapted LDA model for latent activity discovery

Data: spatiotemporal attributes grouped by individual $\boldsymbol{x}$, $\boldsymbol{d}$, $\boldsymbol{t}$, $\boldsymbol{r}$, and $\boldsymbol{m}$
Result: activity assignments $\boldsymbol{z}$, and related latent variables π, θ, ϕ, μ, τ, η, and λ
begin
  randomly initialize $\boldsymbol{z}$, and set up auxiliary variables n_z, u_{mz}, v_{zx}, w_{zd}, s_z, S_z, q_z, and Q_z ;
  foreach *iteration* **do**
    for $i \leftarrow 1$ **to** N **do**
      $z \leftarrow z_i$, $x \leftarrow x_i$, $d \leftarrow d_i$, $t \leftarrow t_i$, $r \leftarrow r_i$, $m \leftarrow m_i$;
      $n_z = n_z - 1$, $u_{mz} = u_{mz} - 1$, $v_{zx} = v_{zx} - 1$, $w_{zd} = w_{zd} - 1$;
      $s_z = s_z - t$, $S_z = S_z - t^2$, $q_z = q_z - \log(r)$, $Q_z = Q_z - \log(r)^2$;
      for $k \leftarrow 1$ **to** Z **do**
        calculate the conditional probability $P(z_i = k|\cdot)$ based on Eq. (13) ;
      end
      $z' \leftarrow$ sample from $P(z_i|\cdot)$;
      $n_{z'} = n_{z'} + 1$, $u_{mz'} = u_{mz'} + 1$, $v_{z'x} = v_{z'x} + 1$, $w_{z'd} = w_{z'd} + 1$;
      $s_{z'} = s_{z'} + t$, $S_{z'} = S_{z'} + t^2$, $q_{z'} = q_{z'} + \log(r)^2$, $Q_{z'} = Q_{z'} + \log(r)^2$;
    end
  end
  for $j \leftarrow 1$ **to** M **do**
    calculate π_j based on Eq. (6) ;
  end
  for $k \leftarrow 1$ **to** Z **do**
    calculate θ_k, ϕ_k, μ_k, τ_k, η_k, and λ_k based on Eqs. (7) to (12) ;
  end
  return $\boldsymbol{z}$, π, θ, ϕ, μ, τ, η, λ ;
end

characterize each individual by fewer activities. Similarly, a sparser θ_z or ϕ_z means that the model prefers to characterize each activity by fewer locations or days of week. In this case, because there are only 7 days of week ($D = 7$), θ_z is unlikely to be sparse, and the choice of γ has little effect on the results. β, on the other hand, determines how "similar" two locations need to be (that is, how often they need to co-occur across different contexts) to find themselves assigned to the same activity. Therefore, for lower values of β, the model is reluctant to assign multiple activities to a given location. However, because of the mixed land use patterns in London, especially around train stations, more than one activity is likely to be accessible from each station. As a result, β may be higher than the choice commonly used for topic modeling in text analysis, e.g., $\beta = 0.1$ in Griffiths and Steyvers (2004). The Dirichlet hyperparameters used in this study are summarized as follows:

- $\alpha_z = 50/Z$, for $z = 1, ..., Z$; this choice is based on Griffiths and Steyvers (2004).
- $\beta_x = 1$, for $x = 1, ..., X$.
- $\gamma_d = 1$, for $d = 1, ..., D$.

3.4.2. Normal-Gamma Priors

The normal-gamma distribution is a bivariate four-parameter family of continuous probability distributions. For arrival time $t \sim \text{Normal}(\mu_z, \tau_z)$ with unknown mean μ_z and precision τ_z, the prior is $\text{NormalGamma}(\mu_0, \kappa_0, \epsilon_0, \tau_0)$. It means that $\tau_z \sim \text{Gamma}(\epsilon_0, \tau_0)$ and $\mu_z \sim \text{Normal}(\mu_0, \kappa_0\tau_z)$. τ_z is determined by the shape parameter ϵ_0 and rate parameter τ_0 of the Gamma distribution. In other words, $E(\tau_z) = \epsilon_0/\tau_0$, $\text{Var}(\tau_z) = \epsilon_0/\tau_0^2$. As τ_z controls the degree of concentration for the distribution of t given activity z, a larger τ_z means that the distribution of t is more concentrated on μ_z. It is preferable to avoid very small τ_z values (i.e., very large variances) so that the model may discover meaningful temporal patterns. One way to achieve this is to set both ϵ_0 and τ_0 very large, as this will reduce $\text{Var}(\tau_z)$ without decreasing $E(\tau_z)$.

On the other hand, μ_z follows a normal distribution with mean μ_0 and variance $1/(\kappa_0\tau_z)$. Therefore, μ_0 should be our guess about where μ_z is, and κ_0 is our certainty about μ_0. Unless there are strong beliefs about μ_z, it is preferable to set μ_0 to the sample average, and κ_0 to a small value so that a larger range of possible values of μ_z can be explored.

For arrival time $r \sim \text{LogNormal}(\eta_z, \lambda_z)$ and its prior $\text{NormalGamma}(\eta_0, \nu_0, \omega_0, \lambda_0)$, the same properties apply. The difference is that the specific hyperparameter values need to chosen with respect to $\log(r)$ instead of r. Both t and r are measured in hours, but λ_z should be larger than τ_z, as the scale of $\log(r)$ is much smaller.

Based on preliminary tests, the following hyperparameter values seem to work well based on the dataset available:

- $\mu_0 = 14, \kappa_0 = 0.01$; 14 is roughly the mean of t in the data.
- $\epsilon_0 = 10^4, \tau_0 = 10^4$; the expected standard deviation of $t|z$ is 1.
- $\eta_0 = 2.5, \nu_0 = 0.01$; $\exp(2.5) = 12$ is roughly the mean of r in the data.
- $\omega_0 = 10^5, \lambda_0 = 10^3$; the expected standard deviation of $\log(r)|z$ is 0.1.

4. Data

To test the proposed model, we use a dataset of pseudonymised trip records from more than 100,000 unique smart cards over two years. The data were made available by Transport for London. We assume each card corresponds to an individual. The public transportation system in London consists of several modes. However, the dataset only covers the rail-based modes, including London Underground, Overground, and part of National Rail. Therefore, the dataset can only capture a subset of the trips taken by each individual, which is typical for large-scale mobility data sources.

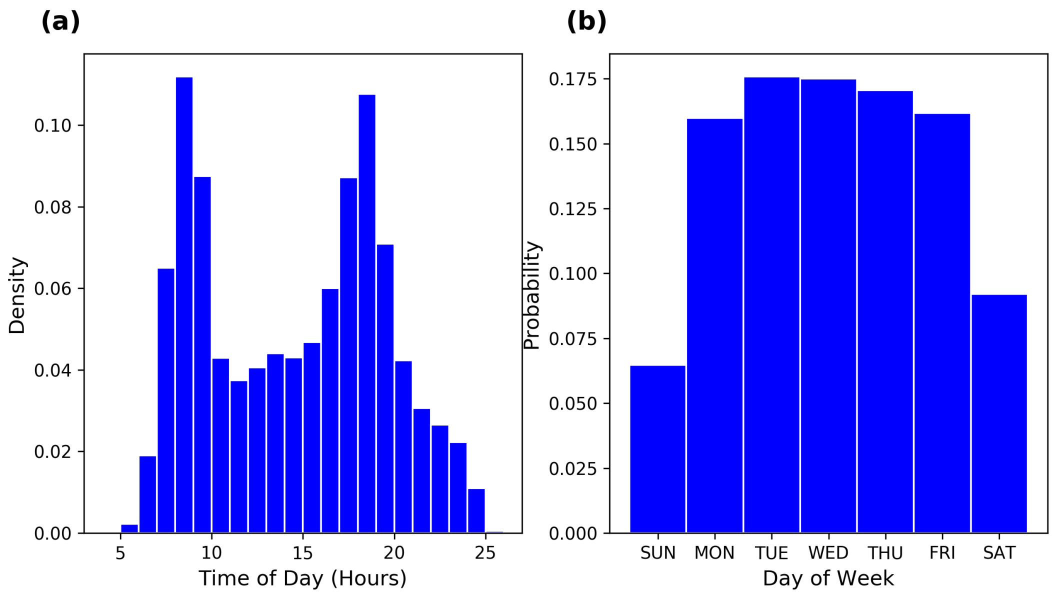


Figure 2: Distribution of arrival time and day of week

For each trip in the dataset, we extract an activity episode with four attributes—location x, day of week d, arrival time t, and duration r. The first three attributes are directly obtained from the smart card transaction recorded when the individual exits the transit system at the destination station. The duration for an activity episode is defined as the difference between the end time of the preceding trip and the start time of the succeeding trip. However, because only a subset of trips are recorded in the data, an individual may make another trip between the two consecutive trips observed in the data. This was referred to as a *hidden visit* in Zhao et al. (2016). In order to determine the location of an activity episode, it is important to ensure that the destination of the preceding trip and the origin of the succeeding trip are close to each other. In this study, for an activity episode to be included in the analysis, the distance between the destination of the preceding trip and the origin of the succeeding trip has to be smaller than a distance threshold $\delta = 2$ km.

Note that this does not guarantee the exclusion of hidden visit. For example, an individual may travel by taxi from location A to location B before returning to A; this can not be observed from the smart card data. In this case, however, the hidden visit to B may be considered as a sub-episode of the activity episode at A. As the duration, or "elapsed time interval" (Zhao et al., 2016), becomes longer, the activity episode is more likely to involve such hidden visits and become less "pure". Therefore, it is important to set a duration threshold. In this study, for an activity episode to be included in the analysis, the difference between the end time of the preceding trip and the start time of the succeeding trip has to be smaller than a duration threshold $T = 72$ hours. The choice of T is to allow the model to identify potential activities related to weekends.

We include only those who have at least 20 observations, i.e., $N_m \geq 20$. After data pre-processing, we obtain 3,339,187 activity episodes from 20,667 individuals. Figure 2 illustrates the distribution of the arrival time and day of week. Figure 2(a) shows the distribution of arrival time t, which is dominated by the morning and afternoon peaks. Figure 2(b) shows the distribution of day of week d; it is clear that there are more trips on weekdays than weekends.

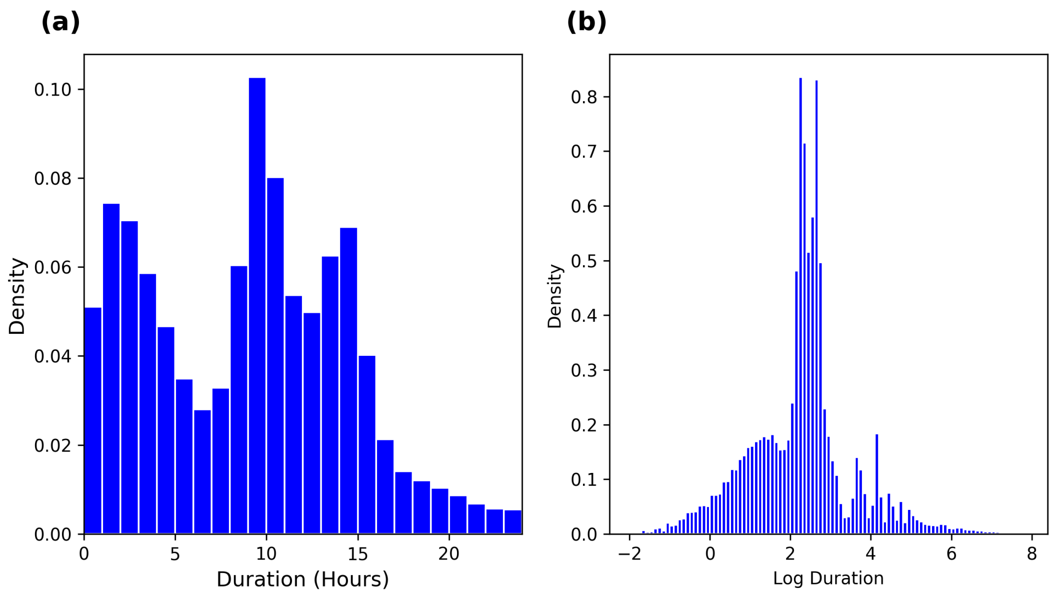


Figure 3: Distribution of duration

The distribution of the duration r is shown in Figure 3, in the original scale on the left, and the log scale on the right. Based on Figure 3(a), r is characterized by three modes—13-15 hours, 9-11 hours, and 1-3 hours. They probably correspond to the three categories of activities—*home*, *work*, and *other*. Figure 3(b) shows the distribution of $\log(r)$ before applying the duration threshold $T = 72$ ($\log(72) = 4.28$). Note that two modes can be seen on the right of the three aforementioned modes, one around 38 hours (1 day + 2 nights), and the other around 63 hours (2 days + 3 nights). This may correspond to people who do not travel for one or two days, most likely over weekends.

Figure 4 presents the top 20 most visited locations (in this case, metro stations) in the data, and their corresponding probabilities. Oxford Circus is by far the most popular

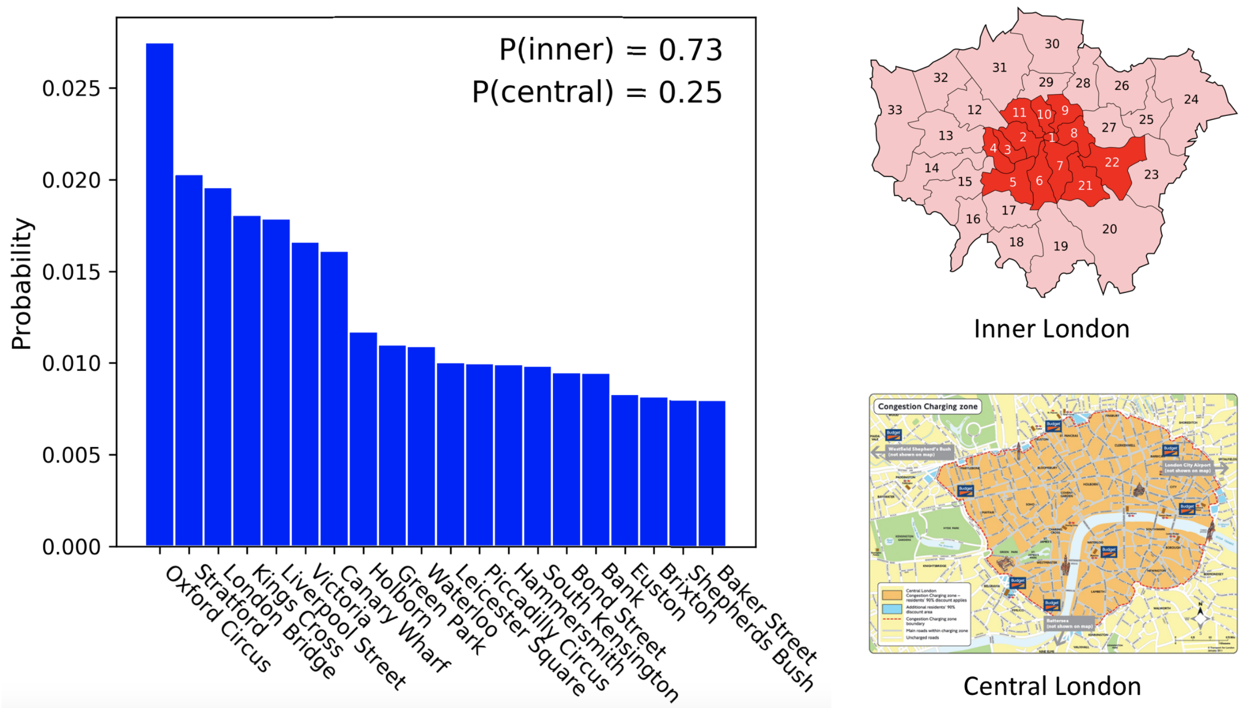


Figure 4: Distribution of locations

destination, followed by Stratford and London Bridge. In total, 665 stations appear in the dataset, i.e., $X = 665$. As one might expect, most stations have low probabilities, and are located in the suburban areas. Showing the top stations may not effectively reflect the overall spatial patterns. Therefore, we use P(inner) to indicate the total probability of all the stations within Inner London, and P(central) for Central London. Inner London refers to the group of London boroughs, and the City of London, which form the interior part of Greater London. The top right map shows all the boroughs of Greater London, with the dark red area referring to Inner London. Central London is located at the core of Inner London. In this study, Central London is defined as the area within the congestion charging zone, which is highlighted in the bottom right map. P(inner) and P(central) are shown in the top right corner of Figure 4. It means that, based on the sample dataset, 73% of the activity episodes occur in Central London and 25% in Inner London.

5. Results

The overall framework of the proposed model introduced in Section 3 is implemented in Python programming language, while the core computational procedure of Gibbs sampling is written in Cython to reduce computational time. The actual time required to estimate the parameters depends on the sample size, the dimensionality of $\boldsymbol{x}$, $\boldsymbol{d}$, $\boldsymbol{t}$, and $\boldsymbol{r}$, as well as the number of activities Z. A typical setup for the data used in this paper took less than 30 min.

Given the data and aforementioned hyperparameters, the number of activities Z still needs to be selected based on the use case. In the literature, perplexity is often used to choose Z (Farrahi and Gatica-Perez, 2011; Hasan and Ukkusuri, 2014). However, the interpretability of the results is also very important. In practice, a smaller number of activities is preferable as it is easier to examine and interpret the results, and less computationally costly to fit the model. A set of potential values of Z are tested: 3, 5, 10, 15, and 20. For exploration purposes, let us start with $Z = 3$.

5.1. Home, Work and Other

Traditionally, the simplest way to categorize activities are to classify them into three basic types: *home*, *work* (including school), and *other*. By setting $Z = 3$, we can test whether the model generate the same activities, as a sanity check.

When $Z = 3$, the summary of the 3 discovered activities is shown in Table 3. The columns of the table indicate the following:

- Index: the ID of the discovered activity
- $E(\pi_{mz}|z)$: the average activity proportion per individual, or $\frac{1}{M}\sum_{m=1}^{M}\pi_m$. Note that the activities are not equally important; some activities are more prevalent than others. To reflect this, the discovered activities are ranked by importance, i.e., the activity index indicates the order of importance for that activity.
- $E(\mu_z)$: the expected μ_z based on its posterior distribution. In the table, the value is converted to clock time format for readability.
- Weekend: the aggregated probability of an activity z starting on weekends. It is computed based on ϕ_z.
- $\exp(E(\eta_z))$: the exponential of expected η_z. It is roughly the mode of the distribution of $r|z$. The unit is an hour.
- P(inner): the aggregate probability of an activity z occurring within inner London. It is computed based on θ_z.
- Description: a short interpretation of the activity. As the model does not explicitly provide a meaningful label for the results, this has to be generated based on the researcher's domain knowledge.

Table 3: Summary of activity characteristics ($Z = 3$)

Index	$E(\pi_{mz}\|z)$	$E(\mu_z)$	Weekend	$\exp(E(\eta_z))$	P(inner)	Description
A3-1	0.44	14:06	0.23	3.70	0.85	Other
A3-2	0.31	19:07	0.14	17.80	0.53	Home
A3-3	0.25	08:30	0.04	9.85	0.86	Work

Figure 5 shows the distributions of $P(t|z)$, $P(d|z)$, $P(r|z)$, and $P(x|z)$ for each activity z. In the figure, each column corresponds to an activity, and each row corresponds to a specific attribute. $P(x|z)$ is shown in the fourth row. Because it is difficult to visually present the probabilities of all 665 locations, we only show the top 10 locations related to each activity. P(inner) and P(central) are embedded in the figure to represent the overall spatial pattern of each activity.

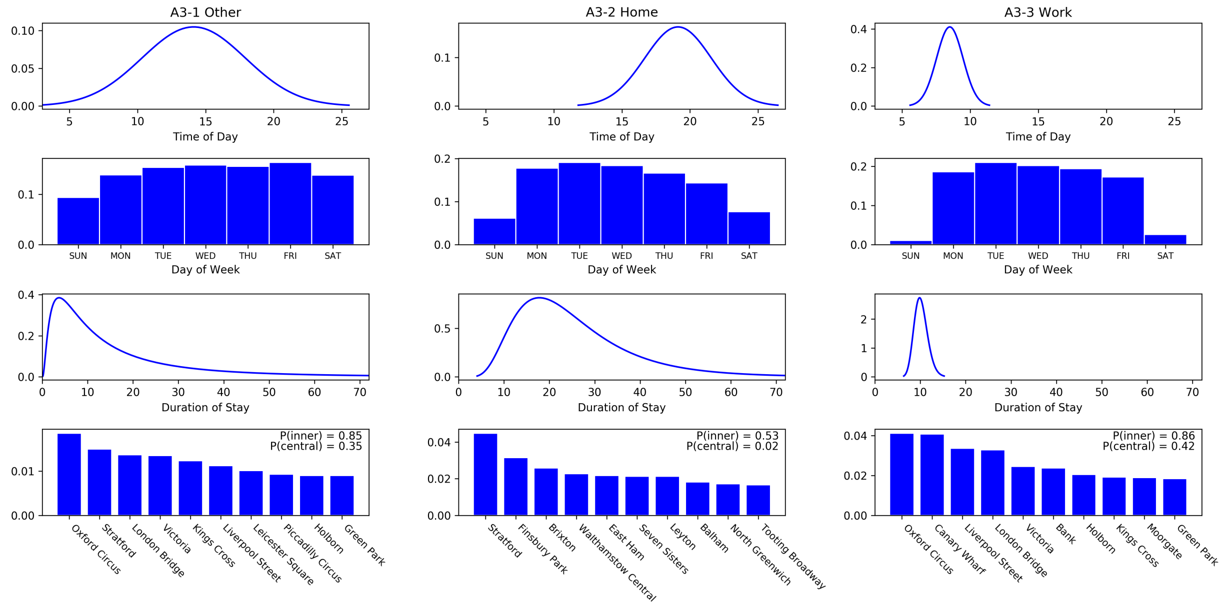
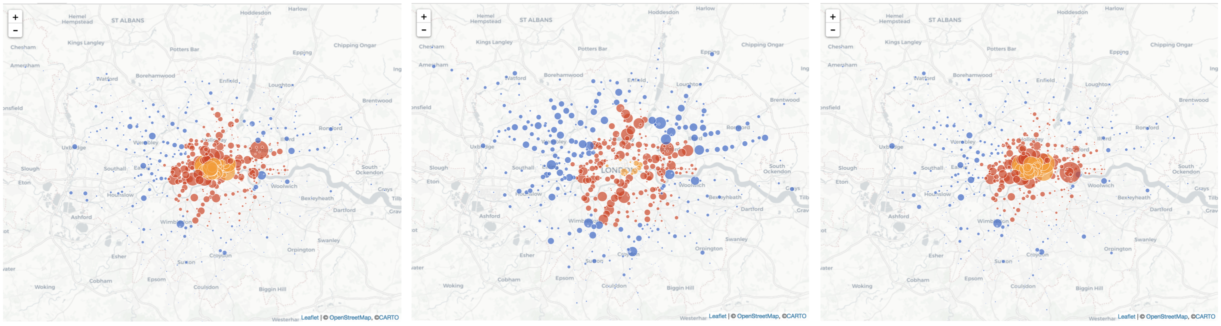


Figure 5: Spatiotemporal distributions by activities ($Z = 3$)

It is relatively easy to identify activities that are related to work or school, as such activities typically start around morning rush hours on weekdays. Based on Table 3 and Figure 5, A3-3 fits this description. Its $P(t|z)$ concentrates around 9 am and its $P(d|z)$ is much higher on weekdays than weekends (96% vs 4%). Some of the most likely locations are important employment centers, such as Canary Wharf and Bank, and the duration is around 10 hours.

In addition, we can identify activities related to home by examining $P(t|z)$ and $P(r|z)$, because people mostly stay home at night, and P(inner) and P(central), because residential locations tend to be more dispersed than other types of locations. A3-2 is a likely candidate. It typically starts at 7 pm and lasts for 18 hours, covering the whole night time. Note that both $P(t|z)$ and $P(r|z)$ are much more spread out for A3-2 than for A3-3. This is not surprising as time spent at home tends to be more flexible than time spent at work/school.

The remaining activity, A3-1, likely includes all other activities, including, but not limited to, errands, meetings, dinners, movies, restaurants, and bars/clubs. They tend to be short in duration, with a mean of less than 4 hours, and may occur at any time of day on any day of week. Both A3-1 and A3-3 have high concentration in Inner London (above 85%). The detailed spatial distributions of the three activities are shown in Figure 6. Each circle in the map indicates a location, with its size proportional to its probability in θ_z. The color is used to represent its centrality—orange means that the location is within Central London, red means within Inner London but outside Central London, and blue means Outer London. Clearly, A3-2 is much more dispersed spatially than the other two activities.



(a) A3-1 (b) A3-2 (c) A3-3

Figure 6: Spatial distributions of A3-1, A3-2, and A3-3

5.2. Model Comparison

With no ground truth activity labels, it is challenging to directly benchmark the model performance in terms of accuracy. Also, for many travel demand modeling tasks, the objective is not always to accurately predict activity labels, but to use activities to explain travel behavior. Therefore, in this section, the comparison is done in terms of how well the activity categorization explains spatiotemporal behavior, measured by the goodness of fit to the data. As a simple validation, we compare our model results against two baseline models adapted from rule-based methods in the literature. The first one (baseline 1) is based on a assumption from Hasan et al. (2013) in which an individual's home and work locations are assumed to be the most visited and second most visited places, respectively. The second (baseline 2) is inspired by Alexander et al. (2015), which determine home and workplaces with the following two rules:

- An individual's home is the place with most visits on weekends and weekdays between 7pm and 8am.
- An individual's work location is the place (not previously labeled as home) to which the individual travels the maximum total distance from home, or $max(d * n)$, where n is the total number of visits to the given place, and d is the its distance to the individual's home location.

In a way, the only difference between the proposed topic model and the baseline models is how z_{mn} is assigned; the former estimates it through Bayesian inference while the latter determine it through simple rules. Once z_{mn} is given, we can calculate the likelihood for either approach. The process to evaluate the goodness of fit of the baseline models is summarized as follows:

1. For each individual $m = 1, 2, ..., M$,
   (a) Use predefined rules to find the home and work locations, denoted as $X_m^{(1)}$ and $X_m^{(2)}$ respectively.
   (b) For each activity episode of the individual $n = 1, 2, ..., N_m$,
      i. If $x_{mn} = X_m^{(1)}$, $z_{mn} = 1$
      ii. If $x_{mn} = X_m^{(2)}$, $z_{mn} = 2$
      iii. Otherwise, $z_{mn} = 3$
2. With $\boldsymbol{z}$ known, calculate π, θ, ϕ, μ, τ, η, and λ based on Eqs. (6) to (12). For comparability, we use the same hyperparameters as discussed in Section 3.4.
3. Calculate the log likelihood and perplexity based on Eqs. (2) to (5).

Table 4 summarizes the goodness of fit metrics of the baseline models and the proposed model with various choice of Z. While baseline 2 fits the data better than baseline 1, neither come close to the proposed model with equal number of activity types ($Z = 3$). This means that the activity categorization discovered the model can better capture the spatiotemporal patterns in the data compared to rule-based activity categorization. This is not surprising, as the model is fitted through learning the representation of the data. As Z increases, the model fit improves.

Table 4: Comparison of model fit

Model	Num of Categories	Log Likelihood	Perplexity
Baseline 1	3	-42734546	361453.77
Baseline 2	3	-42150323	303437.15
Topic Model ($Z = 3$)	3	-37496314	75295.42
Topic Model ($Z = 5$)	5	-36667325	58742.21
Topic Model ($Z = 10$)	10	-36007846	48214.59
Topic Model ($Z = 15$)	15	-35489251	41279.08
Topic Model ($Z = 20$)	20	-34955179	35177.80

Similarly, we can examine the key statistics of the activities determined by the rule-based method, which are shown in Tables 5 and 6. For baseline 1, while it is relatively easy to distinguish *other* due to its shorter duration, higher probability of occuring on weekends and higher concentration in Inner London, the difference between *home* and *work* are not that obvious. This is partly because the simplicity of the rules used, as visit frequency alone may not be able to differentiate between the two types of activities. For baseline 2, the distinction between *home* and *work* is clearer, but not always makes sense. For example, the results show that *home* has far higher concentration in Inner London than *work*, which contradicts the intuition about the urban land use patterns. This is likely caused by the rule that requires the work location to have greatest total distance from home, which might prioritize the locations in the peripheral areas of the city.

In contrast, the discovered activities described in Table 3 are much more distinctive, and their summary statistics arguably more intuitive. As the total variability within the data is constant, the higher distinguishability between groups natually implies lower heterogeneity within groups. This is a desirability quality to have in activity categorization.

Table 5: Summary of activity characteristics for baseline 1

Label	$E(\pi_{mz}\|z)$	$E(\mu_z)$	Weekend	$\exp(E(\eta_z))$	P(inner)
Home	0.34	14:34	0.12	11.11	0.69
Work	0.27	14:06	0.10	10.43	0.71
Other	0.39	14:26	0.22	4.83	0.82

Table 6: Summary of activity characteristics for baseline 2

Label	$E(\pi_{mz}\|z)$	$E(\mu_z)$	Weekend	$\exp(E(\eta_z))$	P(inner)
Home	0.34	14:35	0.12	11.11	0.68
Work	0.16	14:59	0.16	9.02	0.29
Other	0.50	14:16	0.17	6.46	0.80

In travel demand modeling, human activity information is often used to predict travel behavior. Therefore, another way to evaluate model performance is to see how well the discovered activity patterns can predict travel behavior. As an example, we specifically focus on predicting the departure time of the next trip of an individual, which is equivalent to predicting the duration of the current activity episode. It has been shown that the start time of the trip is the least predictable attribute (Zhao et al., 2018b) for next trip prediction. An estimation of the latent activity type (based on location and start time) may help improve prediction performance. To evaluate the predictive performance, we calculate the predictive likelihood of the actual duration r_{mn} for each activity episode, by summing over all possible latent activity types, as shown in Eq. (14). The median of the predictive log likelihoods across all observations is used for model comparison.

$$P(r_{mn} \mid \boldsymbol{z}^{-mn}, \boldsymbol{r}^{-mn}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{t}) = \sum_{z=1}^{Z} P(r_{mn} \mid z_{mn} = z) P(z_{mn} = z \mid \boldsymbol{r}^{-mn}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{t}) \tag{14}$$

where $P(z_{mn} = z \mid \boldsymbol{r}^{-mn}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{t})$ can be calculated in similar fashion as Eq. (13). Note that for heuristic baseline models, this would be deterministic, which means it can only take the value of either 0 or 1.

The model performance is summarized in Table 7. The results show that, compared to the baseline models, the latent activity patterns discovered by the topic model can help us better predict the departure time of the next trip. As Z increases, the prediction performance improves significantly. While a large number of latent activities may limit the interpretability of the results, it could be used to improve the prediction accuracy of travel behavior.

Table 7: Model comparison for predicting the departure time of the next trip

Model	Num of Categories	Predictive Log Likelihood (Median)
Baseline 1	3	-1.046
Baseline 2	3	-1.126
Topic Model ($Z = 3$)	3	-0.970
Topic Model ($Z = 5$)	5	-0.903
Topic Model ($Z = 10$)	10	-0.835
Topic Model ($Z = 15$)	15	-0.730
Topic Model ($Z = 20$)	20	-0.563

5.3. Finding Structure in Activity Patterns

In the proposed model, Z serves as a controller for the level of granularity in the discovered activity patterns. As we increase the value of Z, more specific activity patterns start to emerge. Figure 7 shows how activities evolve as Z increases from 3 to 5, and then to 10. The three groups of activities from left to right represent the corresponding activities discovered when $Z = 3, 5$, and 10, respectively. The specific results are the latter two groups are summarized in Sections Appendix A and Appendix B. The width (or thickness) of the path connecting two activities indicates the number of observations whose activity assignments change from the one on the left to the one on the right when Z increases. The wider the path, the stronger the connection between the two activities.

When Z increases from 3 to 5, the general home activity A3-2 splits into two subcategories—Home (or other) over weekend A5-5, and home between two workdays A5-3 and A5-4, the latter two of which are differentiated based on their spatial patterns (discussed later). This distinction makes sense, as they have very different temporal patterns in both duration and day of week. A5-5 has distinctively longer duration (48 vs 14 hours) and higher concentration on Fridays. This is likely because many commuters do not travel as much during weekends. Another possible reason is that people tend to travel to other cities during weekends, which would explain the high concentration on major train stations (e.g., King's Cross). Also, when Z reaches 10, half-day work A10-10 is also distinguished as a unique pattern, with relatively shorter duration than general work activity A3-3 (6 vs 10 hours). Overall, the work-related activities are relatively isolated because of their inflexible time schedules. Home and other activities are more connected, as both exhibit some long-duration behavior. For example, it is challenging for the model to distinguish between traveling outside London, and staying home over the weekend.

When Z is small, the temporal pattern plays a more important role in differentiating activities. As Z increases, the spatial attribute becomes increasingly significant. In addition to the difference between A5-3 and A5-4, the spaital pattern $P(x|z)$ also explains the difference between A10-3, A10-6, and A10-9, as well as between A10-4, A10-5, A10-7, and A10-8. All of these activities are related to commuting, either going to work or staying at home between workdays. The model's tendency to differentiate commuting-related activities through spatial patterns is driven by the fact that people's home and work locations are

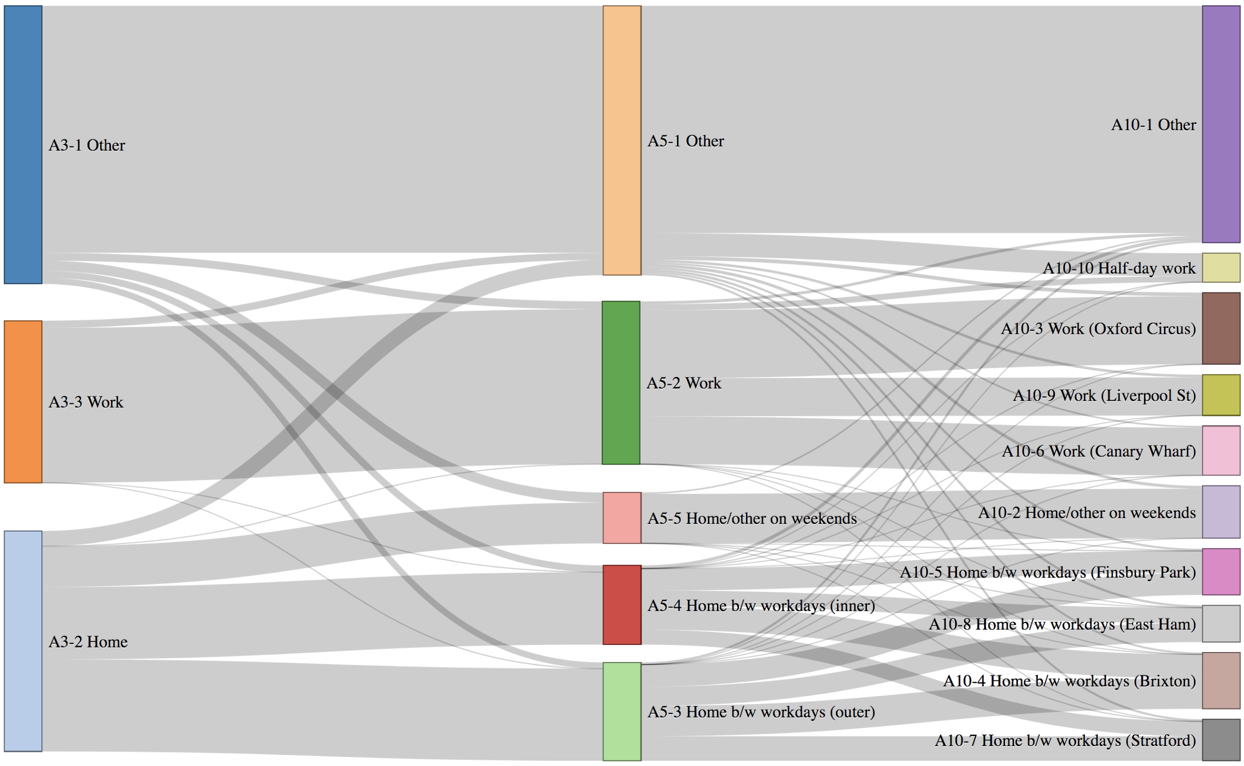


Figure 7: Evolution of discovered activities when $Z = 3, 5, 10$

typically fixed; for most people, there are no interchangeable locations for home or work. As a result, categorizing activities by locations can help explain part of the inter-individual variability, but less so for the intra-individual variability. This is useful for some human mobility tasks where personalization is important, e.g., individual mobility prediction. But if the goal is to study the general time allocation behavior, this might be less helpful. Depending on the application, the balance between temporal and spatial attributes may be adjusted via hyperparameters. For example, a higher β value would reduce the importance of the spatial attribute.

Conventional wisdom tells us that both *home* and *work* are clearly defined and homogeneous activity types, while *other* can be further differentiated into shopping, entertainment, etc. However, the model results show a different story. Although *other* is associated with the largest proportion of observations, the model is reluctant to split it into multiple subgroups when Z increases. This is likely because there is less clear spatiotemporal structure within *other*, compared to *home* and *work*.

In addition to the similarity between activities, we can also examine the co-occurence patterns. This can be done at the individual level. Based on the proposed model, an individual m is characterized by an individual-specific activity distribution π_m. By definition, π_m is a vector of length Z that corresponds to a categorical probability distribution over Z activities; in other words, $\sum_{z=1}^{Z} \pi_{mz} = 1\ \forall m$. Thus π_m can be used as a normalized latent feature vector to describe an individual's activity pattern, or the combination of activities. Correlation may exist between activities. If π_{mj} and π_{mk} are positively correlated across individuals, it means that Activities j and k are more likely to co-occur for the same individual. Figure 8 shows the correlation matrix across the 10 activities discovered by the model when $Z = 10$. Overall, there is no particularly strong correlation between any pair of activities. As expected, positive correlation is found between one of the work-related activities (A10-3, A10-6, A10-9) and one of the home activities (A10-4, A10-5, A10-7, A10-8), which makes sense as it takes two activities to form a commuting pattern. In contrast, the correlation within each group is mostly negative. Again, this is because an individual's home and work locations are fixed.

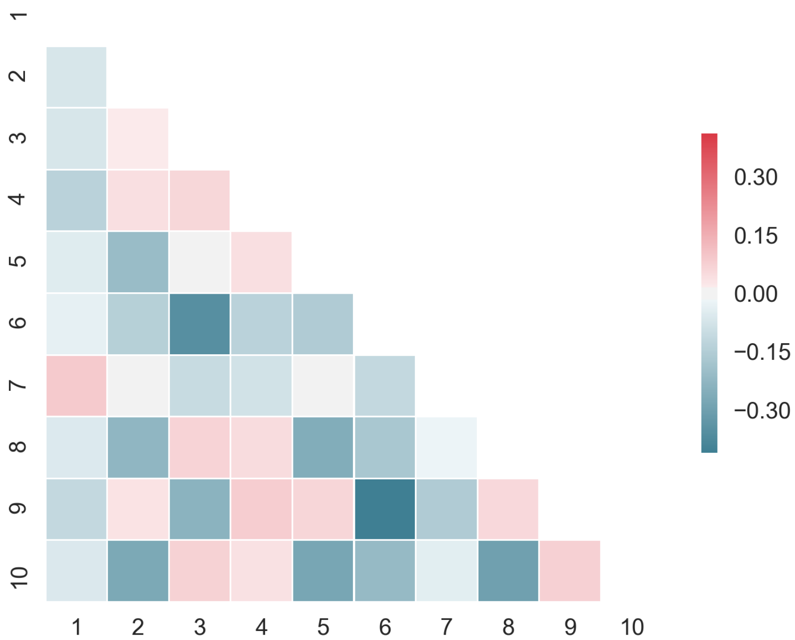


Figure 8: Correlation matrix across activities ($Z = 10$)

6. Discussion

Although automatically collected spatiotemporal records can accurately capture the time and location of human mobility, they do not explicitly provide behavioral semantics underlying the data, e.g., activity types. While many prior works studied *activity inference* (i.e., predicting predefined activity categories), less have focused on *activity discovery* (i.e., finding representative activity categories). In this study, we propose a model to discover latent activities from human mobility data in an unsupervised manner. The proposed model extends the LDA topic model by incorporating multiple heterogeneous dimensions of individual mobility. Specifically, four spatiotemporal attributes—the location, arrival time of day, arrival day of week, and duration of each activity episode—are used in the model to uncover the hidden activity structure, where each "topic" represents a latent activity with a distinct distribution over these attributes. The model is tested with different numbers of activities Z. When $Z = 3$, the model can successfully distinguish the three most basic types of activities—*home*, *work*, and *other*. Compared to rule-based approaches, the proposed model achieves much better goodness of fit. The results also demonstrate how new patterns emerge as Z increases. When Z is small, the temporal pattern plays a more important role in differentiating activities. As Z increases, the spatial attribute becomes increasingly significant. Despite the conventional wisdom that *home* and *work* are more homogeneous than *other*, the model finds more specific subpatterns in *home* and *work*. In addition, positive correlation is found between activities related to work, and activities related to staying home between workdays. The model is general and can be extended for other sources of data where activity episodes are extractable.

This study makes it possible to enrich human mobility data with representative and interpretable activity patterns without relying on predefined activity categories or heuristic rules. On one hand, this can help us uncover new activity patterns or structures that may be helpful to consider in activity-based models. For example, we could distinguish between staying home between workdays or over weekends, or between regular work and half-day work, as they have distinctively different temporal patterns. These finding will then help us refine the existing activity categorization used in activity-travel surveys. On the other hand, when the survey data is not available, we may use the model, instead of simple rules, to generate meaningful activity labels, which can then be used for various human mobility modeling tasks. Trained to differentiate spatiotemporal patterns, the model allows us to account for part of behavioral variability through discovered activity types. An example of this is demonstrated in Section 5.2. Furthermore, the individual-level activity distribution may be used to characterize an individual's activity preferences. It provides a way to transform multidimensional spatiotemporal observations into a normalized latent feature vector, which can be easily adopted for user similarity measurement and cluster analysis. Therefore, the model classifies not only activity episodes, but also individuals.

The methodology presented in this paper has several limitations. First, the model is based on random initialization of activity assignment z_{mn}, and different initialization may lead to somewhat different results. We find that the temporal patterns are relatively stable, but spatial patterns related to commuting (to and from work) are not. As each individual typically has a fixed home/work location, there are a large number of possible ways to divide them into subgroups. Therefore, the spatial characteristics of the commuting-related activities may vary across different model runs. Also, as the spatial proximity between locations are not directly captured in the model, the discovered spatial patterns may not match the underlying geographical areas, limiting our ability to interpret them. Future research should consider incorporating spatial proximity in the model. Second, sequential dependency between trips is important for both activity inference and discovery. Although the model preserves some of the sequential relationship in the data through time and duration variables, it does not explicitly use it as a feature. For example, the probability distribution of the current activity should depend on that of the previous one. The challenge is that adding sequential dependency would add significantly more complexity in model structure. The problem of automatically discovering sequences of activities from data is an ongoing problem, with few good solutions in the literature. Section Appendix C discusses one potential way to add sequential structure to the topic model. Third, some activity types cannot be distinguished based on spatiotemporal patterns alone. For example, the model is not able to differentiate shopping from entertainment. Future work should also explore the possibility of data fusion, by cross referencing other data sources such as surveys, land use, points of interests (POIs), events, and social media posts. This can also help with model selection and validation.

LDA is not the only type of topic models that is adaptable for activity discovery or human mobility modeling in general. Many other types of topic models have been developed over the years to address some of the technical limitations of LDA. Typically, preliminary experiments are needed to choose the number of topics for LDA, which may not be ideal for general applications. Nonparametric methods, such as Hierarchical Dirichlet Process, relaxes this constraint by automatically inferring Z from the data (Teh et al., 2006). Also, dynamic topic models have been developed to analyze the evolution of topics over time (Blei and Lafferty, 2006; Wang and McCallum, 2006), which would be useful for human mobility studies as individual travel patterns can change in the long run (Zhao et al., 2018a). The applicability of these methods should be investigated in the future.

Acknowledgements

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

Appendix A. Model Results with 5 Activities

Table A.8 and Figure A.9 show the summary statistics and spatiotemporal distributions for each of the discovered activities, when $Z = 5$. The top two most common activities among them, A5-1 and A5-2, are very similar to A3-1 and A3-3, respectively. Therefore, they likely represent general other and work activities. This suggests the discovered activity patterns are relatively consistent across different values of Z. Note the decrease in the $E(\pi_{mz}|z)$ for A5-1 and A5-2 are mainly because of the symmetric Dirichlet prior α.

On the other hand, the home-related activities are divided into three subcategories. A5-5 represents activities with long duration. Given its high probability of occurring on Fridays, and low values of P(inner) and P(central), a main reason is that many commuters travel much less frequently by rail over weekends in London. In addition, A5-5 may also include out-of-town trips. Its top 2 most likely locations are King's Cross and Stratford. Both are important transportation hubs, and people may use them as gateways to travel to other cities.

Table A.8: Summary of activity characteristics ($Z = 5$)

Index	$E(\pi_{mz}\|z)$	$E(\mu_z)$	Weekend	$\exp(E(\eta_z))$	P(inner)	Description
A5-1	0.37	14:06	0.23	3.38	0.84	Other
A5-2	0.20	8:30	0.04	9.85	0.86	Work
A5-3	0.16	19:05	0.10	14.30	0.46	Home between work-days (outer)
A5-4	0.14	19:23	0.12	14.27	0.66	Home between work-days (inner)
A5-5	0.13	18:06	0.25	48.06	0.54	Home/other on weekends

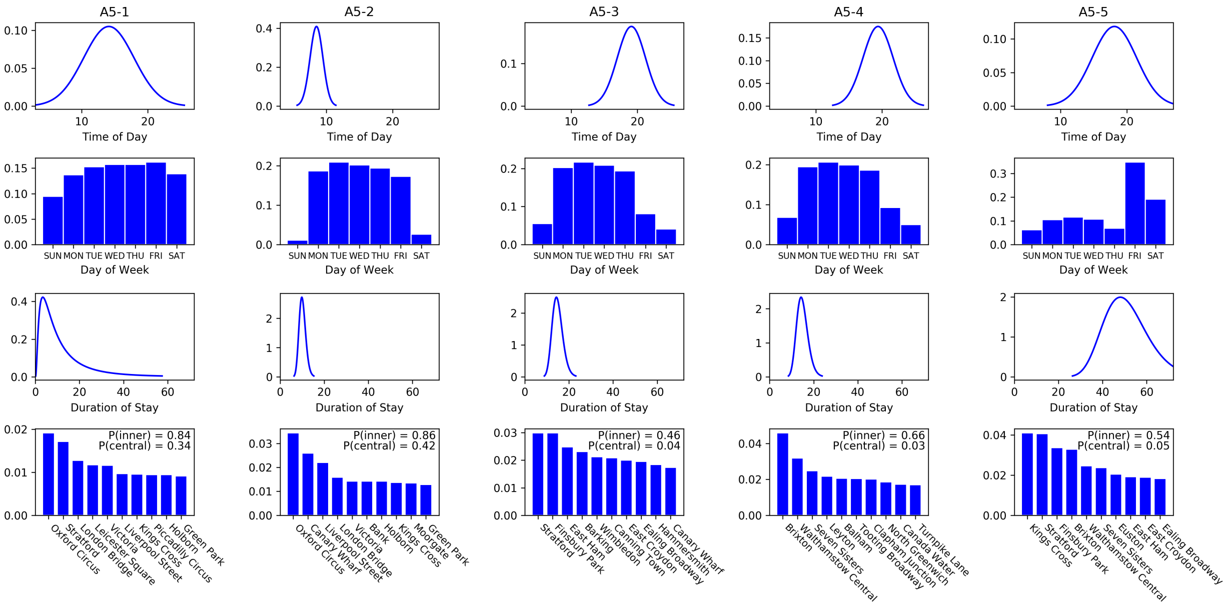


Figure A.9: Spatiotemporal distributions by activities ($Z = 5$)

A5-3 and A5-4 exhibit similar temporal patterns, and are likely associated with the typical afternoon commuting trips, arriving home at around 7:00 pm and stay there for around 14 hours. Interestingly, both have a much lower probability of occurring on Fridays than other weekdays. A possible explanation for this is that most people do not go to work on weekends. As a result, the home activities starting on Friday nights typically have a much longer duration, which is captured by A5-5. The main difference between A5-3 and A5-4 is in their spatial distributions. Note that A5-4 has a relatively higher concentration in inner London, while A5-3 is more dispersed spatially. There is no distinctive geographical boundary that divides the two activities, as the model is oblivious to geographic coordinates of the stations.

Appendix B. Model Results with 10 Activities

Table B.9 and Figure B.10 show the summary statistics and spatiotemporal distributions for each of the discovered activities, when $Z = 10$. Again, some consistent patterns can be identified. A10-1 is similar to A3-1 and A5-1, and A10-2 is similar to A5-5.

Table B.9: Summary of activity characteristics ($Z = 10$)

Index	$E(\pi_{mz}\|z)$	$E(\mu_z)$	Weekend	$\exp(E(\eta_z))$	P(inner)	Description
A10-1	0.30	14:33	0.24	3.02	0.85	Other
A10-2	0.09	17:57	0.25	47.57	0.54	Home/other on weekends
A10-3	0.09	08:34	0.04	9.89	0.90	Work (Oxford Circus)
A10-4	0.08	19:12	0.10	14.31	0.50	Home between workdays (Brixton)
A10-5	0.08	19:09	0.11	14.33	0.60	Home between workdays (Finsbury Park)
A10-6	0.08	08:27	0.08	10.04	0.86	Work (Canary Wharf)
A10-7	0.07	19:06	0.12	14.39	0.48	Home between workdays (Stratford)
A10-8	0.07	19:17	0.12	14.29	0.64	Home between workdays (East Ham)
A10-9	0.07	08:27	0.05	10.08	0.79	Work (Liverpool St)
A10-10	0.07	9:58	0.13	6.09	0.81	Half-day work

A10-3, A10-6, and A10-9 all share similar temporal patterns with A3-3 and A5-2, and thus are all associated with typical work schedules. They mainly differ in $P(x|z)$. A10-10 emerges as a new pattern, whose duration is longer than A10-1 and shorter than A10-3, A10-6, and A10-9. This may represent half-day work shifts or instances when people get off work early. A10-10 also has a higher probability of occurring on weekends, which may indicate that it is associated with atypical work schedules, such as that of a sales person in a shop.

A10-4, A10-5, A10-7, A10-8 all share similar temporal patterns with A5-3 and A5-4, representing staying home over-night between two workdays. All of them have a low probability of occurring on Friday nights. Again, the difference lies in $P(x|z)$. The difference lies in their spatial concentration

Appendix C. Adding Sequentiality to Topic Model

The proposed topic model can be extended to incorporate the sequential structure of human activity-travel behavior. To do this, We could add the sequential dependency either

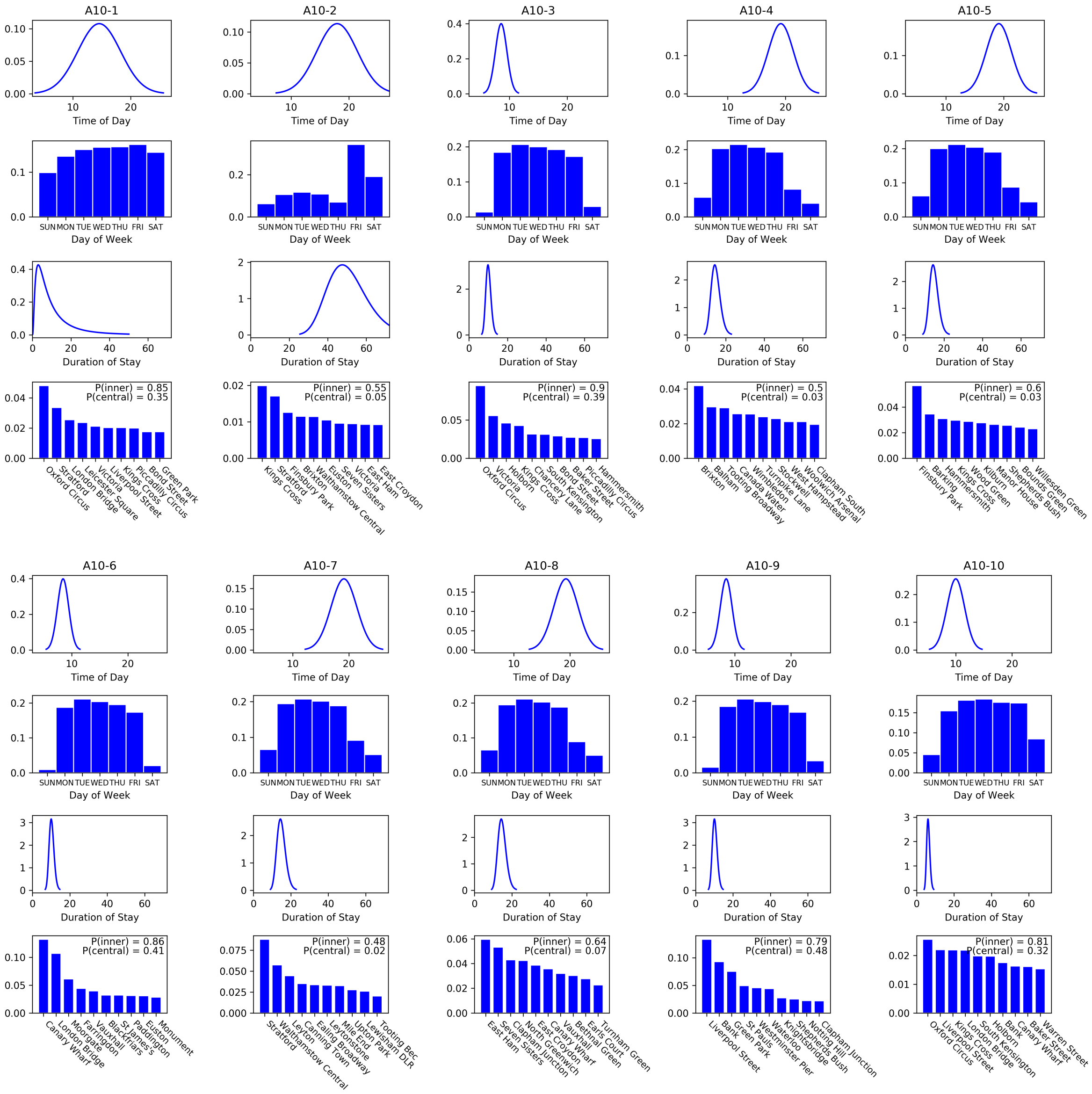


Figure B.10: Spatiotemporal distributions by activities ($Z = 10$)

between activity episodes ($\{x_{mn}, d_{mn}, t_{mn}, r_{mn}\}$), or between latent activity types (z_{mn}). The latter is probably easier as it involves a lower number of dimensions. For simplicity, we only focus on first-order Markovian dependency. For a given individual m, we can illustrate the sequential activity structure in Figure C.11. Note that this resembles an individual-specific Hidden Markov Model (HMM). The difference is that, because of the hierarchical structure of the topic model, some of its parameters can be shared across individuals.

The cost of adding this sequential structure is that it requires the estimation of a Z-by-Z transition matrix for each individual $m = 1, 2, ..., M$, which can be significant when the Z is large. In our dataset, $M = 20667$. If we want to estimate $Z = 10$ latent activities, we

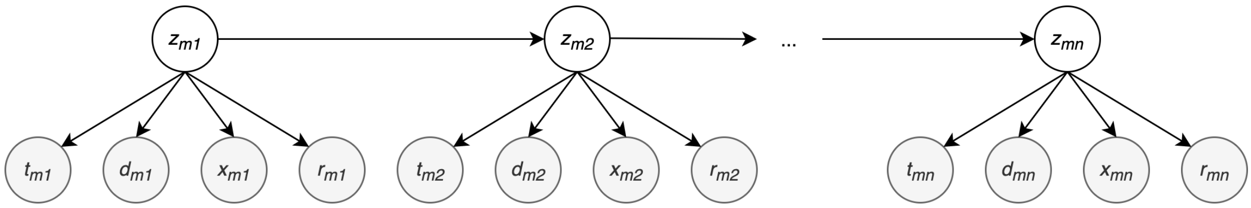


Figure C.11: Illustration of sequential activity structure for individual m

would need to estimate over 2 million additional variables. A much longer observation time period is likely needed. We will reserve it for future research to explore how to estimate this model efficiently and robustly with limited data.

References

Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, *58*, 240–250. URL: http://www.sciencedirect.com/science/article/pii/S0968090X1500073X. doi:10.1016/j.trc.2015.02.018.

Allahviranloo, M., and Recker, W. (2013). Daily activity pattern recognition by using support vector machines with multiple classes. *Transportation Research Part B: Methodological*, *58*, 16–43. URL: http://www.sciencedirect.com/science/article/pii/S0191261513001689. doi:10.1016/j.trb.2013.09.008.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., and Hickman, M. (2018). Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, *87*, 123–137. URL: http://www.sciencedirect.com/science/article/pii/S0968090X17303777. doi:10.1016/j.trc.2017.12.016.

Axhausen, K. W., and Gärling, T. (1992). Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport Reviews*, *12*, 323–341. URL: https://doi.org/10.1080/01441649208716826. doi:10.1080/01441649208716826.

Bhat, C. R., and Koppelman, F. S. (1999). Activity-Based Modeling of Travel Demand. In *Handbook of Transportation Science* International Series in Operations Research & Management Science (pp. 35–61). Springer, Boston, MA. URL: https://link.springer.com/chapter/10.1007/978-1-4615-5203-1_3. doi:10.1007/978-1-4615-5203-1_3.

Blei, D. M., and Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning* ICML '06 (pp. 113–120). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1143844.1143859. doi:10.1145/1143844.1143859.

Blei, D. M., and McAuliffe, J. D. (2010). Supervised Topic Models. *arXiv:1003.0783 [stat]*, . URL: http://arxiv.org/abs/1003.0783. ArXiv: 1003.0783.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. URL: http://www.jmlr.org/papers/v3/blei03a.html.

Bowman, J. L., and Ben-Akiva, M. E. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, *35*, 1–28. URL: http://www.sciencedirect.com/science/article/pii/S0965856499000439. doi:10.1016/S0965-8564(99)00043-9.

Chen, Z., Zhang, Y., Wu, C., and Ran, B. (2019). Understanding Individualization Driving States via Latent Dirichlet Allocation Model. *IEEE Intelligent Transportation Systems Magazine*, *11*, 41–53. doi:10.1109/MITS.2019.2903525.

Côme, E., Randriamanamihaga, A., Oukhellou, L., and Aknin, P. (2014). Spatio-temporal Analysis of

Dynamic Origin-Destination Data Using Latent Dirichlet Allocation. Application to Vélib' Bikesharing System of Paris. URL: https://trid.trb.org/view/1287424.

Eagle, N., and Pentland, A. S. (2009). Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, *63*, 1057–1066. URL: http://link.springer.com/article/10.1007/s00265-009-0739-0. doi:10.1007/s00265-009-0739-0.

Farrahi, K., and Gatica-Perez, D. (2009). Learning and Predicting Multimodal Daily Life Patterns from Cell Phones. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* ICMI-MLMI '09 (pp. 277–280). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1647314.1647373. doi:10.1145/1647314.1647373.

Farrahi, K., and Gatica-Perez, D. (2011). Discovering Routines from Large-scale Human Locations Using Probabilistic Topic Models. *ACM Trans. Intell. Syst. Technol.*, *2*, 3:1–3:27. URL: http://doi.acm.org/10.1145/1889681.1889684. doi:10.1145/1889681.1889684.

Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*, 5228–5235. URL: http://www.pnas.org/content/101/suppl_1/5228. doi:10.1073/pnas.0307752101.

Han, G., and Sohn, K. (2016). Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. *Transportation Research Part B: Methodological*, *83*, 121–135. URL: http://www.sciencedirect.com/science/article/pii/S0191261515002593. doi:10.1016/j.trb.2015.11.015.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., and González, M. C. (2013). Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics*, *151*, 304–318. URL: https://doi.org/10.1007/s10955-012-0645-0. doi:10.1007/s10955-012-0645-0.

Hasan, S., and Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, *44*, 363–381. URL: http://www.sciencedirect.com/science/article/pii/S0968090X14000928. doi:10.1016/j.trc.2014.04.003.

Hidayatullah, A. F., and Ma'arif, M. R. (2017). Road traffic topic modeling on Twitter using latent dirichlet allocation. In *2017 International Conference on Sustainable Information Engineering and Technology (SIET)* (pp. 47–52). doi:10.1109/SIET.2017.8304107.

Huynh, T., Fritz, M., and Schiele, B. (2008). Discovery of Activity Patterns Using Topic Models. In *Proceedings of the 10th International Conference on Ubiquitous Computing* UbiComp '08 (pp. 10–19). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1409635.1409638. doi:10.1145/1409635.1409638.

Kim, Y., Pereira, F. C., Zhao, F., Ghorpade, A., Zegras, P. C., and Ben-Akiva, M. (2014). Activity Recognition for a Smartphone Based Travel Survey Based on Cross-User History Data. In *2014 22nd International Conference on Pattern Recognition* (pp. 432–437). doi:10.1109/ICPR.2014.83.

Kusakabe, T., and Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, *46*, 179–191. URL: http://www.sciencedirect.com/science/article/pii/S0968090X14001612. doi:10.1016/j.trc.2014.05.012.

Lee, S. G., and Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, *6*, 1–20. URL: https://link.springer.com/article/10.1007/s12469-013-0077-5. doi:10.1007/s12469-013-0077-5.

Liao, L., Fox, D., and Kautz, H. (2005). Location-based Activity Recognition Using Relational Markov Networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* IJCAI'05 (pp. 773–778). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. URL: http://dl.acm.org/citation.cfm?id=1642293.1642417.

Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368/. doi:10.1186/s40064-016-3252-8.

McNally, M. G. (2007). The Four-Step Model. In *Handbook of Transport Modelling* (pp. 35–53). Emerald Group Publishing Limited volume 1 of *Handbooks in Transport*. URL: http://www.emeraldinsight.

com/doi/abs/10.1108/9780857245670-003. doi:10.1108/9780857245670-003.

Montoliu, R. (2012). Discovering Mobility Patterns on Bicycle-Based Public Transportation System by Using Probabilistic Topic Models. In P. Novais, K. Hallenborg, D. I. Tapia, and J. M. C. Rodríguez (Eds.), *Ambient Intelligence - Software and Applications* Advances in Intelligent and Soft Computing (pp. 145–153). Springer Berlin Heidelberg.

Peng, C., Jin, X., Wong, K.-C., Shi, M., and Liò, P. (2012). Collective Human Mobility Pattern from Taxi Trips in Urban Area. *PLOS ONE*, *7*, e34487. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0034487. doi:10.1371/journal.pone.0034487.

Rasiwasia, N., and Vasconcelos, N. (2013). Latent Dirichlet Allocation Models for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 2665–2679. doi:10.1109/TPAMI.2013.69.

Rasouli, S., and Timmermans, H. (2014). Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences*, *18*, 31–60. URL: https://doi.org/10.1080/12265934.2013.835118. doi:10.1080/12265934.2013.835118.

Sun, L., and Axhausen, K. W. (2016). Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transportation Research Part B: Methodological*, *91*, 511–524. URL: http://www.sciencedirect.com/science/article/pii/S0191261516300261. doi:10.1016/j.trb.2016.06.011.

Sun, L., Chen, X., He, Z., and Miranda-Moreno, L. F. (2019). Pattern Discovery and Anomaly Detection of Individual Travel Behavior using License Plate Recognition Data. URL: https://trid.trb.org/view/1572444.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*, 1566–1581. URL: http://dx.doi.org/10.1198/016214506000000302. doi:10.1198/016214506000000302.

Wang, X., and McCallum, A. (2006). Topics over Time: A non-Markov Continuous-time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '06 (pp. 424–433). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1150402.1150450. doi:10.1145/1150402.1150450 event-place: Philadelphia, PA, USA.

Zhao, Z., Koutsopoulos, H. N., and Zhao, J. (2018a). Detecting pattern changes in individual travel behavior: A Bayesian approach. *Transportation Research Part B: Methodological*, *112*, 73–88. URL: http://www.sciencedirect.com/science/article/pii/S0191261518300651. doi:10.1016/j.trb.2018.03.017.

Zhao, Z., Koutsopoulos, H. N., and Zhao, J. (2018b). Individual mobility prediction using transit smart card data. *Transportation Research Part C: Emerging Technologies*, *89*, 19–34. URL: http://www.sciencedirect.com/science/article/pii/S0968090X18300676. doi:10.1016/j.trc.2018.01.022.

Zhao, Z., Zhao, J., and 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.

Zheng, J., Liu, S., and Ni, L. M. (2014). Effective Mobile Context Pattern Discovery via Adapted Hierarchical Dirichlet Processes. In *2014 IEEE 15th International Conference on Mobile Data Management* (pp. 146–155). volume 1. doi:10.1109/MDM.2014.24.

Zou, Q., Yao, X., Zhao, P., Wei, H., and Ren, H. (2018). Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. *Transportation*, *45*, 919–944. URL: https://doi.org/10.1007/s11116-016-9756-9. doi:10.1007/s11116-016-9756-9.