Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Unveiling mobility patterns beyond home/work activities: A topic modeling approach using transit smart card and land-use data

Nima Aminpour, Saeid Saidi [*]

Department of Civil Engineering, University of Calgary, AB T2N 1N4 Calgary, Alberta, Canada,

ARTICLE INFO

Keywords:
Activity pattern
Topic modeling
Transit smart card
Land use data
COVID-19 impact

ABSTRACT

In this paper, a probabilistic topic modeling algorithm called Latent Dirichlet Allocation (LDA) is implemented to infer trip purposes from activity attributes revealed from smart card transit data in an unsupervised manner. While most literature focused on finding patterns for home and work activities, we further investigated non-home and non-work-related activities to detect patterns associated with them. Temporal attributes of activities are extracted from trip information recorded by Tehran subway's automatic fare collection system. In addition, land-use data is also incorporated to further enhance spatial attributes for non-home/work activities. Various activity attributes such as start time, duration, and frequency in addition to land-use data are used to infer the activity purposes and patterns. We identified 14 different patterns related to non-commuting activities on the basis of both their temporal and spatial attributes including educational, recreational, commercial, and health and other service-related activity types. We investigated passengers' activity pattern and behavior changes before and during COVID-19 pandemic by comparing the discovered patterns. For recreational patterns it is revealed that not only has the number of recreational patterns dropped, but also the duration of recreational activities decreased. Morning patterns of educational activities have also been eliminated and number of commercial activities was decreased during COVID-19. The proposed model demonstrates the ability to capture travel behavior changes for different disruptions using smart card transit data without performing costly and time consuming manual surveys which can be useful for authorties and decision makers.

1. Introduction

Understanding trip purposes plays a crucial role in unraveling the intricate relationship between human activities and travel behaviors. This knowledge serves as a cornerstone for effective urban planning, transportation strategies, and policy formulation. For instance, Sato and Maruyama (2020) utilized trip purposes to model departure times in travel surveys, while Mo et al. (2022) employed trip purpose information to predict individuals' next trips. Traditionally, trip purpose information has been gathered through surveys and interviews, providing insights into the motivations and intentions driving people's movements. However, the limitations of these methods, including small sample sizes and recall biases, have underscored the need for more robust and automated approaches.

The advent of automatically collection data technologies like transit smart card systems has revolutionized the fare collections, also allowing for automatic and non-invasive tracking of travel patterns on a larger scale. These systems provide details about trip timings, destinations, and durations, enabling a deeper understanding of the dynamics between passenger activities and travel behaviors. They reveal a lot of spatial and temporal information about transit trips including inferencing origin–destination demand, travel demand analysis, monitor and control the transit system, crowding management specially in near capacity situations. Despite the wealth of data obtainable from transit smart cards, understanding the motivations behind each trip remains a challenge. Travel purposes are multifaceted, influenced by individual choices, economic factors, and diverse range of other variables. Moreover, a single trip often serves multiple purposes, such as commuting to work and running errands en-route. To address this challenge, recent research has turned to machine learning to uncover the hidden motivations behind travel using smart card data. These methods leverage patterns and trends in the data to discern why people travel. By employing both supervised and unsupervised learning techniques, models can accurately predict travel purposes based on various factors such as time, location, and mode of transportation.

In this paper, we introduce a machine learning approach to

[*] Corresponding author.
E-mail addresses: nima.aminpour@ucalgary.ca (N. Aminpour), ssaidi@ucalgary.ca (S. Saidi).

https://doi.org/10.1016/j.tbs.2024.100905
Received 29 September 2023; Received in revised form 26 August 2024; Accepted 6 September 2024
Available online 20 September 2024

determine the motivations behind travel using transit smart card data. Specifically, we propose a methodology that combines machine learning with Bayesian reasoning to infer the purpose of each trip (i.e. activity type performed after the trip). Our focus extends beyond the typical home and work activities. By adding the land-use data, we can consider a wider variety of activities such as educational, commercial, recreational. As such, we can have a more comprehensive picture of why people travel and how to expect changes to travel pattern given different disruptions. For instance, we investigate how the COVID-19 pandemic has altered activity patterns, akin to the study conducted by Lin et al. (2023) on the pandemic's impact on urban visitor travel patterns, illustrating our method's capability.

In the subsequent sections of this paper, we first review the literature of using data driven approach in inferring the travel purpose or activity pattern, then we discuss the methodology employed for inferring travel purposes using transit smart card and land-use data. In the results section, we explain the outcome of our discoveries, including the inferred travel purposes across various activities. We then present the case of comparing activity pattern changes during COVID-19 pandemic using the model developed. Finally, the conclusion will summarize the main takeaways and contributions of this research.

2. Literature review

The trip purpose inferencing literature is divided into three major groups namely supervised, unsupervised, and rule-based methods. In rule-based approaches (Anda et al., 2017; Devillaine et al., 2012) passenger travel behavior is quantitatively analyzed by features or thresholds associated with known activity categories. These features and thresholds are specified by researchers based on domain knowledge to distinguish trip purposes. Kuhlman (2015) and Lee and Hickman (2014) used rule-based processing to infer trip purpose by considering four major attributes of activity duration, departure time, frequency, and card type for the trip purpose inference. Alsger et al. (2018) Inferred trip purposes based on destination location and temporal features. Household survey data were also used for validation of their result. Many studies applied different rulesets for identifying home trips, such as what proposed by Zou et al. (2018), that the first and last tap-in stations of a fare payment cardholder may be considered as their potential home locations. Alexander et al. (2015) found that passengers' homes are the most visited locations between 7 pm and 8 am on each day of the week. In accordance with Hasan and Ukkusuri (2014), passengers' homes and workplaces are the first and second most visited locations as trips to home locations are more frequent than trips to work locations. Rules and thresholds are also derived for identifying work trips, such as (Chakirov and Erath, 2012) which found that the subsequent trip is typically observed around six hours or more after the work trip; Devillaine et al. (2012) found work trips in Gatineau, Canada take over five hours. Travelling in the morning and returning in the evening peak was also considered as a work trip for employer-based smart cards (Lee and Hickman, 2014). Rule-based approaches, however, have limitations. They may not universally apply to all circumstances, as regional differences in working hours can lead to unique work patterns that cannot be deduced using criteria from another region. Also, distinguishing short acitivites and transfers is a challanging problem (Nassir et al., 2015). Additionally, unforeseen events like the COVID-19 pandemic can disrupt established rules due to significant changes in mobility patterns.

With the advancement of information technologies, data fusion is becoming more feasible, resulting in the use of supervised learning methods as an alternate to the rule-based approach in the trip-purpose discovery. Using supervised approaches, the first step is to train samples with labeled information using data fusion approaches and then use the trained model to predict trip purposes based on trip attributes from automatic fare collection (AFC) or GPS data. Literature in this area differs more in their modeling approaches. As an example, A naive Bayes probabilistic model was developed by Kusakabe and Asakura (2014) and Kuhlman (2015) estimating trip purpose by using probabilistic purpose inference models. Faroqi and Mesbah (2021) proposed clustering passengers approach based on their trip temporal attributes and the trip purposes using passenger trip sequences. Kim et al. (2021) developed a random forest model trained on household travel survey data using spatiotemporal attributes extracted from smart card data and geographic information. Sari Aslam et al. (2021) developed an artificial neural network (ANN) trained on smart card and Points of Interest (POI) data, along with trip purposes obtained from a survey to predict the purpose of trips under primary activities (e.g. home and work) as well as secondary activities (e.g. entertainment, eating, shopping, dropping off/picking up children). Social media data and land-use data were used to construct a multi-objective Convolutional Neural Network (CNN) to determine social demographic parameters and mobility characteristics of passengers (Zhang et al., 2020). Montini et al. (2014) developed a random forest model for trip purpose classification using GPS tracks and accelerometer data. Trip purposes estimation with a logit allocation model approach using GPS and household travel survey data was proposed by Chen et al. (2010).

While supervised learning can help with better validation of the results, supervised approaches in trip purpose research pose challenges in acquiring reliable labeled data due to the difficulties and costs associated with data labeling. To address this issue, unsupervised algorithms have been developed, allowing researchers to uncover trip purposes without relying on pre-existing labeled datasets. One common approach is clustering passengers based on their travel behaviors using temporal and spatial attributes. Agard (2009) and Ma et al. (2013) inferred trip purposes by analyzing passengers' travel behaviors and clustering them based on attributes such as start time, frequency, and activity location. However, these approaches have limitations and may not capture the nuances associated with different activity types. To address these limitations, Allahviranloo and Recker (2013) used a multiclass Support Vector Machine (SVM) to separate passenger travel data into groups and infer activity types. Their approach outperformed traditional multinomial logit models. Eagle and Pentland (2009) employed Principle Component Analysis (PCA) to create eigenbehavior vectors and cluster individuals based on their behaviors. Peng et al. (2012) applied nonnegative matrix factorization to analyze passenger traffic patterns based on latent patterns in taxi data. Incorporating spatial information in trip analysis has also been explored. Han and Sohn, (2016) and Mo et al. (2022) both utilized hidden Markov models in their respective studies, with the former employing a continuous hidden Markov model (CHMM) for imputing activity sequences using AFC datasets and land use information, and the latter using a hidden Markov model to explore activity patterns using smart card data. Yu et al. (2021) identified metro passenger mobility patterns by embedding spatiotemporal data derived from AFC and POIs into low-dimensional trip vectors using a stacked auto-encoder.

In broader terms, these techniques do not work effectively when dealing with data that is grouped, where multiple trips linked to the same individual show a high correlation. Since activity patterns differ from passenger to passenger, it's essential to factor in the distinct behaviors exhibited by each individual. A hierarchical structure then should be adopted, which is capable of capturing variations both between individuals and within an individual at various hierarchy levels (Zhao et al., 2020c). To address the issue of discovering activity patterns in datasets with correlated trips of the same individuals, Latent Dirichlet Allocation (LDA) is a popular approach that has been successfully applied in natural language processing (NLP) and has shown promising results in discovering hidden topics in text data. LDA is one of the natural language processing algorithms which was introduced by Blei et al. (2003). LDA is a probabilistic model originally designed for discrete data, such as text. It represents each item in a collection as a mixture of topics, and each topic as a mixture of topic probabilities. LDA approach has been implemented in different areas such as image processing (Rasiwasia and Vasconcelos, 2013), text classification (Blei et al., 2003), and bioinformatic area (Liu et al., 2016). In transportation domian, researchers implement LDA approach for driver behavioral analysis (Chen et al., 2019) and extract bike sharing patterns (Come et al., 2014). Activity discovery is another area that researchers recently implemented LDA algorithm. An unsupervised probabilistic topic model, adapted from LDA, is proposed by Zhao et al. (2020c) to discover interpretable trip purpose categories from individual-level spatiotemporal data. It is demonstrated that the proposed methodology can effectively describe the temporal and spatial attributes of work-related and home-related activities using London smart card data. Wang et al. (2016) implemented the LDA algorithm to generate users' trip topics who have trips to commercial districts. Users are then clustered into groups with different trip purposes based on the eigenvectors extracted from the user topic distribution. The latest mentioned studies have focused more on temporal characteristics than spatial information, resulting in difficult-to-interpret results. This challenge was overcome by using spatially related data sets, such as POIs and land-use datasets. Li et al. (2021) proposed a LDA model for identifying the trip purposes for the remaining trips, after finding the home/work trip purposes using heuristic rules. The attribute considered in their study were arrival time, age group, stay duration, and the point of interest tag for the destination.

Despite attempts to address the heterogeneous behavior in literature using algorithms such as LDA, current methods utilized for activity inference studies still fall short in providing detailed trip purpose classifications. Initially, researchers focused on inferring basic commuting trips such as work and home (Zhao et al., 2020b; Zou et al. 2018; Yu et al. 2021; Sari Aslam et al., 2019), along with educational activities (Chu and Chapleau, 2010; Devillaine et al., 2012; Lee and Hickman, 2014). However, these methods lacked the ability to capture nuances associated with different types of activities. While more recent research has improved upon this limitation by accounting for additional purposes like shopping and recreational activities (Anda et al., 2017; Sari Aslam et al., 2021; Xiao et al., 2016), there is still room for further refinement in order to better infer and classify various activity types accurately by using LDA algorithm that has a hierarchical structure capturing variations both between individuals and within an individual level. Leveraging land-use data with subway smart card data to produce topic groups that represent activity purposes in much greater detail than simply home or work-related activities alone, could be an important new contribution. This approach can provide deeper insights into activity patterns and allow for more targeted interventions for the transport authority based on specific travel patterns and preferences.

3. Methodology

This section presents a comprehensive methodology for analyzing Automated Fare Collection (AFC) data to gain deeper insights into avtivity patterns using a subway system. Our methodology incorporates spatiotemporal and land-use attributes extracted from the AFC and land-use datasets, enabling a detailed analysis of individual travel behavior. By combining the identification of major activity groups and the inference of detailed activity categories, we aim to understand the diverse range of activities individuals engage in between their trips using the public transit system.

3.1. Data preparation

This section focuses on the data used in the study and the necessary data preparation steps. The study utilizes smart card data, which is collected through AFC in subway systems. The AFC data includes temporal information (time, date, day of the week) and spatial information (station location) for each tap of a smart card at origin (tap-in) and destination (tap-out) stations.

Each trip in the AFC data is considered to be associated with an activity performed by an individual. To extract activity attributes, we rely on the information from the trips that occurred before and after each activity. The start time of an activity is determined as the end time (tap-out) of the preceding trip, and the duration of the activity is calculated as the time difference between the end time of the preceding trip and the start time of the succeeding trip.

Spatial attributes, such as the location of the activity, are inferred based on the destination station of the preceding trip and the origin station of the succeeding trip. To ensure accurate location assignment, two rational rules are defined to exclude activities involving other modes of transportation. Firstly, we consider an activity episode only when the distance between the destination of the preceding trip and the origin of the succeeding trip is no more than 2 km. This ensures that the identified activities are confined to a localized area and do not involve other modes of transportation. Secondly, we exclude activities with a duration longer than 20 hours, as they may indicate trips taken using alternative transportation modes or represent infrequent home activities. By adhering to these rules, we can more confidently analyze and interpret mobility patterns related to our specific objectives.

The study also incorporates activity frequency as a feature. For calculating the frequency of similar trips, activities are clustered based on start time and duration. Similar activities within each cluster, occurring at the same station or nearby (within a 2 km radius), are considered repeated activities. The frequency of an activity is then determined by counting the total number of similar activities recorded for an individual. In addition to the AFC data, a land-use dataset is utilized to further analyze activities. The dataset provides information about 25 different land-use categories in the dataset. To simplify the analysis, the detailed land-use categories are grouped into broader categories, including Educational, recreational, commercial, residential, office spaces, and health services.

In order to incorporate land-use data, it is assumed that trips ending at a station are associated with activities performed within a 1 km radius of that station. For each trip ending at a station, we assign a share of the total covered area to each land-use category. Additionally, to make the area shares comparable across all activities, a box-cox transformation is applied to the area share of each category across all activities. The Box-Cox transformation is selected to reduce skewness and stabilize the variance across the data, thereby approximating a more Gaussian-like distribution with a mean of zero and unit standard deviation. This scaling is particularly beneficial when comparing area shares for each land-use category across different stations. The lambda parameter, which is a critical component of the Box-Cox transformation, was determined for each land-use category through maximum likelihood estimation. This approach allowed us to identify the lambda that best stabilized the variance and normalized the distribution of area shares for each category. The implementation of this method yielded values ranging from −3 to +3 for each category's area share. For instance, a positive value indicates that the area share for a specific category is above the average area share of that category across all activities. Conversely, a negative value suggests that the area share is below the average. By utilizing these standardized values, we can effectively identify significant variations in land-use attributes across different activity patterns.

The integration of smart card data and land-use data offers a powerful and comprehensive analytical framework for exploring passenger activities beyond the traditional home and work categories. By simultaneously examining temporal and spatial attributes, the proposed methodology yields valuable insights into individuals' travel patterns and preferences.

3.2. Proposed framework

As discussed earlier, important contributions are made in developing activity inferencing models for home and work-based activities. However, detailed activity inferencing for non-commuting trips are generally neglected in the literature. To address this issue, we formulate the problem as a two-step process (See Fig. 1). The first step focuses on

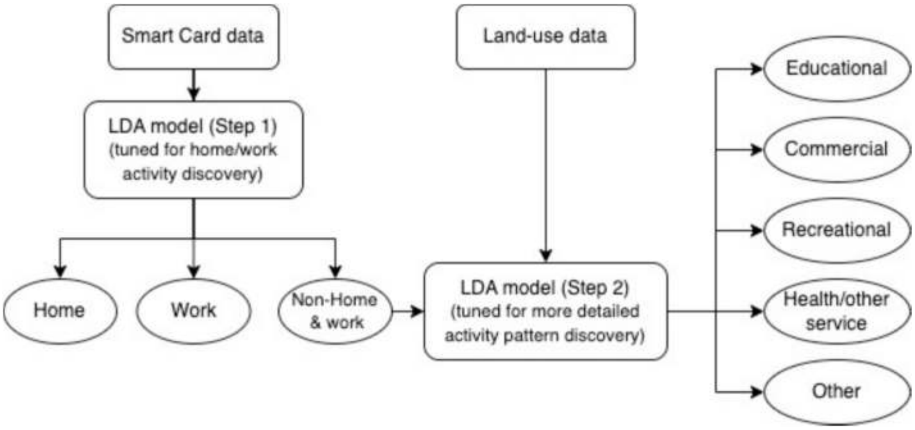

Fig. 1. Research overview flowchart.

identifying major activity groups, including home, work, and non-home & work activities. This initial classification provides a preliminary understanding of general patterns in individuals' travel behavior. By discerning these major activity groups, we can observe the distribution of trips among these categories, identify peak travel periods, and analyze the duration of different activity types.

The second step involves inferring more detailed activity categories related to educational, commercial, recreational, and health & other service-related activities. To accomplish this, we extend our analysis by incorporating land-use data as spatial attributes. By leveraging the AFC dataset together with land-use information, we aim to obtain a more detailed understanding of the various activities passengers engage in at different locations.

3.3. Latent Dirichlet allocation

To identify possible activity categories from the AFC dataset, we utilize a probabilistic topic modeling known as LDA. The model assumes that every document (such as an individual's trip data) comprises multiple topics (akin to activity categories). These topics are characterized by distribution of words, which in the context of this paper, represent spatiotemporal attributes. The LDA framework allows us to learn both the topic mixture of each individual's trip records and the words associated with each topic, facilitating the interpretation of activity categories. We use a similar approach proposed by Zhao et al. (2020a) by employing LDA method but also adding new variables such as frequency and land-use attribute to discover more detailed activity clusters other than home and work-related activities only.

In activity-based analysis, travel behavior is modeled as a series of activities, where each activity has a unique temporal and spatial features. The set of continuous variables in our activity-based analysis framework shown in Fig. 2 includes the start time, activity duration, activity frequency, and land-use area share, and their distribution is dependent on the activity category z_{mn}. We assume that each of these variables, conditioned on z_{mn}, follows a normal distribution with a mean parameter of μ_z and a precision parameter of τ_z specific to each variable. The choice for normal distribution over alternatives was made to be consistent with the symmetric distribution observed across all activity behaviors, particularly regarding start time, frequency and duration. Additionally, the activity type variable for each individual is represented as a categorical distribution with a parameter denoted as π_m, which signifies the distribution of activity types for individual m. The objective of this approach is to estimate the distribution parameters of each attribute (start-time, duration, frequency and land-use area share) for each activity type.

- 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 $z_{nm} \sim \text{Categorical}(\pi_m)$

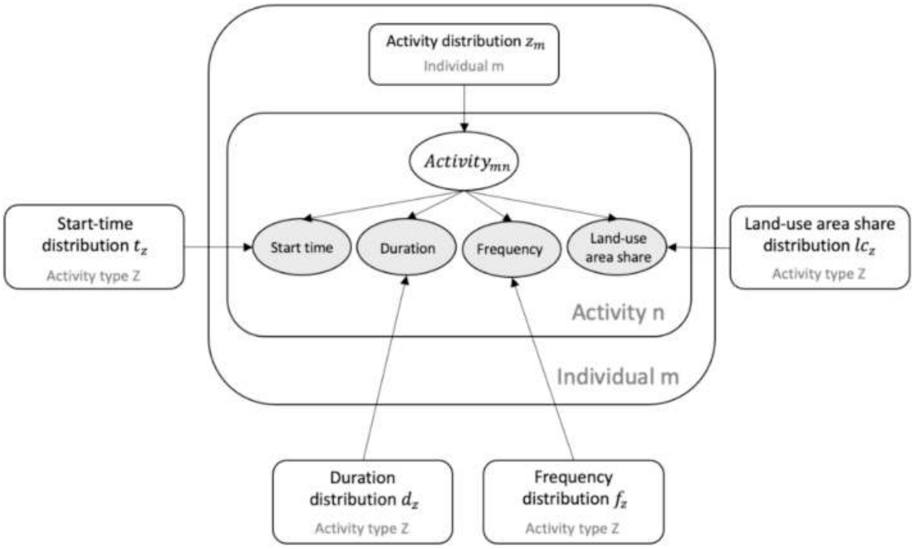

Fig. 2. LDA model considerations for activity variables.

Table 1
Notation explanation.

Notation	Explanation
M	Number of individuals
N	Number of observations
Z	Number of activities
m	Individual number
n	Observation number
N_m	Number of observations of individual m
l_c	Land-use area share for land-use category c
z_{mn}	Activity assignment of observation m of individual m
t_{mn}	Start time of observation n of individual m
d_{mn}	duration of observation n of individual m
f_{mn}	frequency of observation n of individual m (in one month)
$l_{c\,mn}$	Land-use area share of category c of observation n of individual m
π_m	Activity type distribution for individual m
μ_{xz}, τ_{xz}	Mean and precision of attribute x for activity type z
α	Dirichlet hyperparameter for π_m
$\mu_{x0}, \kappa_{x0}, \varepsilon_{x0}, \tau_{x0}$	Normal-gamma hyperparameter for μ_{xz}, τ_{xz}
n_z	Number of observations with activity type z
u_{mz}	Number of observations of individual m with activity type z
s_{xz}	Sum of attribute x for observation with activity type z
S_{xz}	Sum of square of attribute x for observation with activity type z

ii. Sample a time of day $t_{nm} \sim \text{Normal}(\mu_{tz_{mn}}, \tau_{tz_{mn}})$
iii. Sample a duration $d_{nm} \sim \text{Normal}(\mu_{dz_{mn}}, \tau_{dz_{mn}})$
iv. Sample a frequency $f_{nm} \sim \text{Normal}\left(\mu_{fz_{mn}}, \tau_{fz_{mn}}\right)$
v. Sample an area share for land-use type c $lc_m \sim \text{Normal}\left(\mu_{l_{cz\,mn}}, \tau_{l_{cz\,mn}}\right)$

We estimate the distribution parameters based on the available data, which is typically done by utilizing Bayesian inference and conjugate priors. Bayesian inference is a statistical technique that allows for updating the parameter estimates based on new observations. Conjugate priors are a special class of prior distributions that are chosen to ensure that the posterior distribution of the parameter is in the same family as the prior distribution. To estimate the mean and variance parameters of the normal distributions for each attribute, we assume that these parameters are samples from the normal gamma distribution. The normal gamma distribution is a conjugate prior for the normal distribution, which means that we can obtain a closed-form solution for the posterior distribution of the parameters. For the individual activity distribution, we assume that the prior distribution of the parameter π, which describes the distribution of activity types for an individual, follows a Dirichlet distribution. The Dirichlet distribution is a conjugate prior for the categorical distribution, which is the distribution of discrete variables of individual activity type. These prior distributions have hyperparameters assigned by using a combination of expert knowledge and empirical evidence. In particular, the proposed model works under on the assumption that parameters of distributions are generated through the following process. The explanation of all notations is available in Table 1.

- For each activity type z, the conjugate priors can be written as follows:

a) Sample a time of day distribution $\mu_{tz}, \tau_{tz} \sim$ Normal Gamma$(\mu_{t0}, \kappa_{t0}, \varepsilon_{t0}, \tau_{t0})$
b) Sample a duration distribution $\mu_{dz}, \tau_{dz} \sim$ Normal Gamma$(\mu_{d0}, \kappa_{d0}, \varepsilon_{d0}, \tau_{d0})$
c) Sample a frequency distribution $\mu_{fz}, \tau_{fz} \sim$ NormalGamma $\left(\mu_{f0}, \kappa_{f0}, \varepsilon_{f0}, \tau_{f0}\right)$
d) Sample a land-use area share distribution for land-use category c $(l_c)\mu_{lcz}, \tau_{lcz} \sim \text{NormalGamma}(\mu_{l_c0}, \kappa_{l_c0}, \varepsilon_{l_c0}, \tau_{l_c0})$

By considering these assumptions for the prior distribution of each variable, we can proceed to the next step, which explains the approach used to estimate the parameters of distributions for each variable. More comprehensive discussion on the selection of hyperparameters can be found in Aminpour (2024).

3.4. Distribution estimation using Gibbs sampling

To estimate the distribution parameters (such as means and variances) of the spatiotemporal variables for each inferred activity category, we employ the Gibbs sampling method (Griffiths and Steyvers 2004). Gibbs sampling is a Markov Chain Monte Carlo (MCMC) technique that iteratively samples from conditional distributions to approximate the joint distribution. In our case, Gibbs sampling allows us to estimate the distributions of activity-specific variables, such as activity start time, duration, and frequency.

As part of Gibbs sampling process, each observation is randomly assigned an activity category. Then, in each iteration, the probability of assigning all the other categories to that observation is computed using their conjugate priors, eventually the activity category with the greatest probability is assigned to the observation. The new activity assignment is then used to update the distribution parameters. This iterative process continues until convergence is reached, indicating that the estimated distribution parameters are optimized to capture the characteristics of each activity category. The pseudocode of the explained approach is shown in Table 2.

Gibbs sampling algorithm calculates the probability of assigning different activity categories to an observation and eventually assigns the category with the highest probability to that observation. The joint probability of $P(\mathrm{z}, \mathrm{t}, \mathrm{d}, \mathrm{f}, \mathrm{l_c})$ is the probability of assigning an activity category z to observation with a start time of t, activity duration of d and frequency of f and area share l_c, for each land-use type (c) which can be written as $P(\mathrm{z} = z, \mathrm{t} = t, \mathrm{d} = d, \mathrm{f} = f, \mathrm{l_c} = l_c)$. Using the probability property of prior distribution, the joint probability can be expanded as shown in Eq. (1):

$$P(\mathrm{z}=z,\mathrm{t}=t,\mathrm{d}=d,\mathrm{f}=f,\mathrm{l_c}=l_c)= \int_{\pi_m}\int_{\mu_t,\tau_t}\int_{\mu_d,\tau_d}\int_{\mu_f,\tau_f}\prod_{c=1}^{c=C}\int_{\mu_{l_c},\tau_{l_c}} \begin{array}{c} P(\pi_m).P(\mu_{tz},\tau_{tz}).P(\mu_{dz},\tau_{dz}).P\left(\mu_{fz},\tau_{fz}\right).P(\mu_{l_cz},\tau_{l_cz}) \\ .P\left(z,t,d,f,l_c|\pi_m,\mu_{tz},\tau_{tz},\mu_{dz},\tau_{dz},\mu_{fz},\tau_{fz},\mu_{l_cz},\tau_{l_cz}\right) \end{array} \tag{1}$$

Table 2
Implemented LDA model for latent activity discovery.

```
Begin

• Randomly initialize z for each observation          N: total number of observation
• Foreach iteration do:                    Z: total number of activity type
  o For i → 1 to N do:                     z: activity type for each observation
        ▪ For k → 1 to Z do:
            • Calculate the conditional probability of assigning z = k to the observation i
        ▪ Assign the k with maximum probability to the observation I
        ▪ Update the dataset
• Return attributes distribution for each activity type.
End
```

Eq. (1) could be separated as follow:

$$P(\mathrm{z}=z,\mathrm{t}=t,\mathrm{d}=d,\mathrm{f}=f,\mathrm{l_c}=l_c\,) \\ =\left(\int_{\pi_m}P(z|\pi_m)P(\pi_m)\right).\left(\int_{\upmu_\mathrm{t},\tau_\mathrm{t}}P(t|\upmu_\mathrm{tz},\tau_\mathrm{tz})P(\upmu_\mathrm{t},\tau_\mathrm{t})\right).\left(\int_{\upmu_\mathrm{d},\tau_\mathrm{d}}P(d|\upmu_\mathrm{dz},\tau_\mathrm{dz})P(\upmu_\mathrm{d},\tau_\mathrm{d})\right).\left(\int_{\upmu_\mathrm{f},\tau_\mathrm{f}}P\left(f|\upmu_\mathrm{fz},\tau_\mathrm{fz}\right)P\left(\upmu_\mathrm{f},\tau_\mathrm{f}\right)\right).\prod_{c=1}^{c=C}\left(\int_{\upmu_\mathrm{lc},\tau_\mathrm{lc}}P(\mathrm{l_c}|\upmu_{l_c\mathrm{z}},\tau_{l_c\mathrm{z}})P(\upmu_{l_c},\tau_{l_c})\right) \tag{2}$$

By considering the probability property of variables with a prior distribution, each variable's marginal distribution in Eq. (2) can be expanded as Eq. (3):

$$P(\mathrm{z}=z,\mathrm{t}=t,\mathrm{d}=d,\mathrm{f}=f,\mathrm{l_c}=l_c\,)=\frac{u_{mz}+\alpha_z}{\sum_{k=1}^{Z}u_{mk}+\alpha_k} \\ .\mathscr{T}\left(t|2\varepsilon_{t0}+n,\frac{s_{tz}+\kappa_{t0}\mu_{t0}}{n_{tz}+\kappa_{t0}},\frac{\left(\tau_0+\frac{n_zS_{tz}-s_{tz}^2}{2n_z}+\frac{\kappa_{t0}(s_{tz}-n_z\mu t_0)^2}{2n_z(\kappa_{t0}+n_z)}\right)(\kappa_{t0}+n_z)}{\left(\frac{\varepsilon_0+n_z}{2}\right)(\kappa_{t0}+n_z)}\right) \\ .\mathscr{T}\left(\mathrm{d}|2\varepsilon_{d0}+n,\frac{s_{dz}+\kappa_{d0}\mu_{d0}}{n_z+\kappa_{d0}},\frac{\left(\tau_{d0}+\frac{n_zS_{dz}-s_{dz}^2}{2n_z}+\frac{\kappa_{d0}(s_{dz}-n_z\mu_{d0})^2}{2n_z(\kappa_{d0}+n_z)}\right)(\kappa_{d0}+n_{dz})}{\left(\frac{\varepsilon_{d0}+n_z}{2}\right)(\kappa_{d0}+n_z)}\right) \\ .\mathscr{T}\left(f|2\varepsilon_{f0}+n,\frac{s_{fz}+\kappa_{f0}\mu_{f0}}{n_z+\kappa_{f0}},\frac{\left(\tau_{f0}+\frac{n_zS_{fz}-s_{fz}^2}{2n_z}+\frac{\kappa_{f0}\left(s_{fz}-n_z\mu_{f0}\right)^2}{2n_z(\kappa_{f0}+n_z)}\right)(\kappa_{f0}+n_z)}{\left(\frac{\varepsilon_{f0}+n_z}{2}\right)(\kappa_{f0}+n_z)}\right) \\ .\prod_{c=1}^{C}\mathscr{T}\left(\mathrm{l_c}|2\varepsilon_{c0}+n,\frac{s_{cz}+\kappa_{c0}\mu_{c0}}{n_z+\kappa_{c0}},\frac{\left(\tau_{c0}+\frac{n_zS_{cz}-s_{cz}^2}{2n_z}+\frac{\kappa_{c0}(s_{cz}-n_z\mu_{c0})^2}{2n_z(\kappa_{c0}+n_z)}\right)(\kappa_{c0}+n_z)}{(\varepsilon_{c0}+n_z/2)(\kappa_{c0}+n_z)}\right) \tag{3}$$

Before running the Gibbs sampling algorithm, it is important to determine the number of activity categories for which we want to estimate the distribution. One approach is to compare the performance of different models with varying numbers of activity categories and select the model that achieves the best results. Perplexity is commonly used as a measure to evaluate model performance. It quantifies how well a probabilistic model fits a set of observations and can be calculated using the likelihood function. Perplexity is computed by taking the exponent of the negative log likelihood function (Eq. (5)) divided by the total number of observations in the dataset (Eq. (4)).

$$Perplexity = exp\left(-\frac{\log(\mathscr{L}(\mathscr{M}))}{N}\right) \tag{4}$$

$$\mathscr{L}(\mathscr{M}) = P(x,d,f,l_c\,|\mathscr{M}) = \prod_{m=1}^{M}\prod_{n=1}^{N}\sum_{z_{mn}=1}^{Z}P(z_{mn},t_{mn},d_{mn},f_{mn},lc_{mn}) \tag{5}$$

A lower perplexity value indicates a better fit of the model to the data. The likelihood function, used to calculate perplexity, is determined by the chosen number of activity categories and the distribution parameters for each feature. To select the optimal number of activity categories, perplexity can be calculated for different numbers of categories, and the model with the lowest perplexity is chosen. However, it is crucial to consider the interpretability of the results, as selecting too few or too many categories may lead to less meaningful or overly complex output.

In this section, we discussed our methodology to unveil the intrinsic distribution of these attributes across diverse activity categories through the application of the LDA model. To achieve this, the Gibbs sampling technique is employed, enabling the approximation of probability distributions of each attribute. Through iterative updates of activity type assignments for each activity record and their observed attributes, we converge towards the most likely activity type assignments for each record, systematically revealing underlying patterns.

4. Results

4.1. Case study

The case study utilized Tehran AFC data of January 2020 and April 2020, consisting of about 25 million trip records from approximately 2.6 million individuals. To ensure privacy, all individual identifiers within the dataset were anonymized before the data was used for this study. From this extensive dataset, we randomly selected a sample of 10,000 users. This selection process was designed to create a representative subset of the total passenger base, enabling us to draw meaningful inferences about activity patterns while managing the data's scale for analysis effectively. The AFC data included spatiotemporal information on trip origins, destinations, and tap times. Land-use data sourced from Tehran Municipal Studies Organization, was incorporated to capture the surrounding environment of transit stations, with 7 major land-use categories. The area share of each land-use category within subway station coverage was calculated for analysis as mentioned in Section 3.2.

Fig. 3 illustrates all 133 subway stations and 7 subway lines in Tehran. The figure also presents a sample of a station showing the land-use distribution within a 1 km radius of the selected station categorizing the surrounding area into various land uses. The large parcel in red color shows Tehran University which has education land-use.

We carefully considered the spacing between stations, particularly in densely populated areas with a high concentration of stations, to minimize overlap in stations' catchment areas.

4.2. Home/work related activities

The algorithm employed in this study aimed to identify and categorize home and work-related activities based on their distinct temporal patterns in month of January 2020. By considering start time, duration, and frequency attributes as explained in Section 3.1, the algorithm separated these two primary activity types from the remaining activities. The decision to exclude the land-use attribute in this step was deliberate, as incorporating it could introduce unnecessary complexity and potentially obscure the temporal patterns that are crucial for distinguishing home and work-related activities.

Running the algorithm with three major activity groups enabled the discovery of distinct clusters representing home-related, work-related, and other types of activities. Fig. 4 illustrates the distribution of

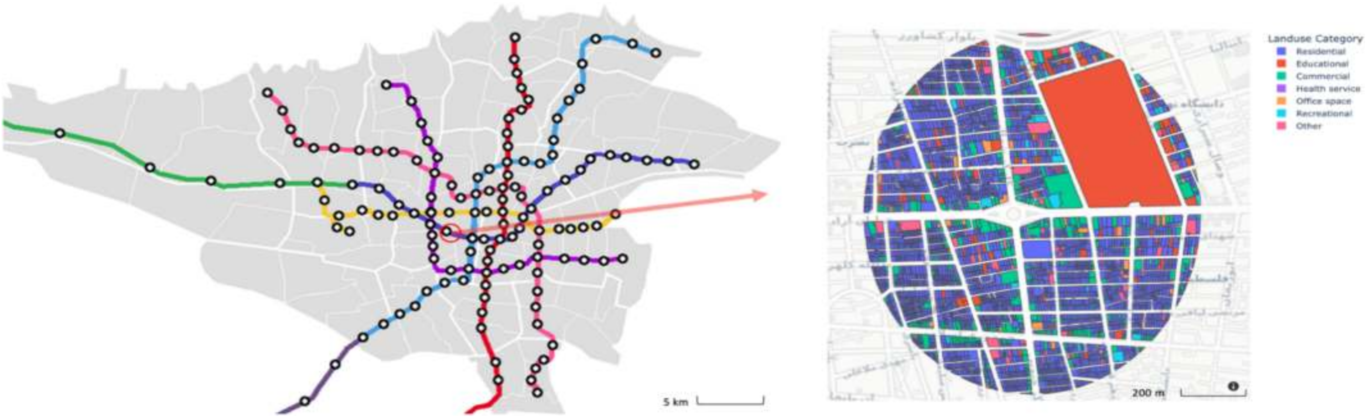

Fig. 3. Subway map of Tehran along with land-use area share for a highlighted station.

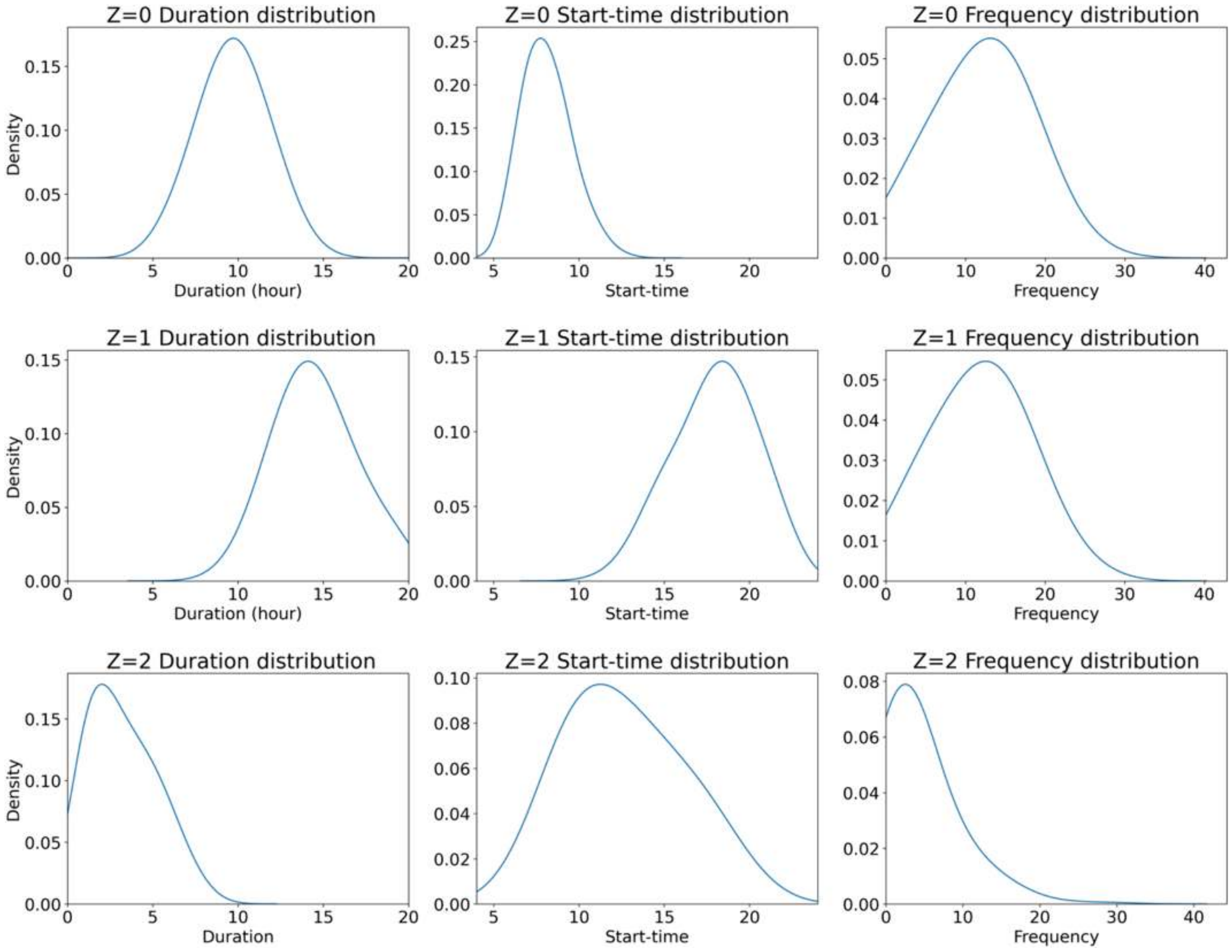

Fig. 4. Attributes distribution for discovered clusters.

Table 3
Summary of activity attributes in discovered clusters.

Activity cluster	Ave. duration (hour)	Ave. start-time (time)	Ave. frequency (in one month)	Activity interpretation	Total share in total number of activity
Z = 0	9.8 h	8	15.8	Work	40 %
Z = 1	14.3 h	18	15.2	Home	42 %
Z = 2	3.2 h	13	1.4	Non-Home&Work	18 %

attributes for each discovered cluster, while Table 3 provides a summary of key activity attributes. Within Cluster Z = 0, primarily encompassing work-related activities, signifying a typical morning commute. This time window spans from 5 a.m. to 12 p.m., peaking at 8 a.m. The duration of these activities ranges from 5 to 15 h, with an average duration of 9.8 h. The activity frequency varies between 1 and 30 occurrences per month, with an average of 16 times. Notably, the distribution places greater emphasis on the left side of the average, indicating a prevalence of more frequent passengers who utilize the subway system predominantly on specific weekdays rather than throughout the entire week. This could be implication of a recurrent pattern where certain individuals consistently use the subway for a limited number of weekdays, repeating the same behavior each week.

Cluster Z = 1 is distinctly linked with home activities and showcases distinctive temporal characteristics. The activities within this cluster primarily commence between 1 pm and late in the night, with the average initiation time centering around 6 pm. These activities tend to extend for an average duration of approximately 14 h, ranging from 9 to 20 h, predominantly occurring during evening time hours. The distribution of their frequency mimics the pattern observed in the work-related cluster, with a notable concentration on the left side of the distribution. This points to a higher occurrence of less frequent subway users who do not utilize the system daily. Considering these insights, the interpretation of these home-related activities is consistent with the identified pattern.

Cluster Z = 2 represents non-home & work types of activities, encompassing a wide range of start times and durations. This cluster includes diverse activity patterns that do not align closely with either home or work-related activities. The activities within this cluster occur infrequently, from 1 to 10 times per month with an average of around 2 times per month, and exhibit significant variation in terms of start times and durations. In light of these observations, we infer this cluster as and non-home & work-related activity which is explored for more detailed activity pattern in next section.

4.3. Inferring non-home and work activities

As discussed in Section 3.2, to further explore and categorize non-home and work activities, the LDA algorithm was run again, focusing specifically on activities identified as “non-home & work” in the previous LDA run. This time, both the land-use information and other activity attributes, including start time, duration, and frequency were taken into account.

The determination of the optimal number of activity clusters was based on the model perplexity which is discussed in detail in Section 3.4. It was observed that as the number of clusters increased, the model’s perplexity decreased, suggesting a more accurate representation of the underlying patterns in the data. However, in order to strike a balance between interpretability and model performance, 14 clusters were chosen as the optimum number for the activity classification.

Fig. 4 provides a comprehensive overview of the discovered activity clusters. Each cluster is described through four key attributes. The first column presents the distribution of start times, revealing the temporal patterns of activities within each cluster. The second column illustrates the distribution of activity durations, revealing information on the time spent on different activities. The third column displays the average standardized value of land-use area share for each land-use category within the cluster. This standardization process allows for a meaningful comparison of the association between specific land-use types and activity clusters. By examining the standardized values, it is possible to identify the land-use categories that exhibit a stronger correlation with certain activity patterns. Lastly, the fourth column depicts the distribution of activity frequencies within a one-month period, providing insights into the occurrence and frequency of different activities. These temporal and spatial attribute distributions were instrumental in inferring activity types for each cluster. In the following paragraphs, we describe how these temporal and spatial attribute distributions were used to infer activity types for each cluster.

Clusters z_b = 5, 7, and 12 (See Fig. 9) are identified as being associated with educational activities, with positive standardized land-use area shares for educational facilities. Clusterz_b = 5 represents frequent, long-duration activities mainly starting in the morning, indicating regular school or university activities. Cluster z_b = 12 also represents frequent, long-duration activities but with a broader range of start times, potentially related to school and university activities starting later in the day. Cluster z_b = 7 represents short-duration, infrequent activities at various hours, such as visits to musical/language institutions, bookstores, and cultural centers near educational institutions.

Clusters z_b = 8 and 13 (See Fig. 6) are inferred to represent commercial activities, occurring in stations with positive standardized commercial land-use area shares. Cluster z_b = 8 represents infrequent, long-duration morning activities, while cluster z_b = 13 represents infrequent, short-duration evening activities.

Clusters z_b = 0, 2, 9, 10, and 11 (See Fig. 7) are associated with recreational activities, occurring more frequently in stations with positive standardized recreational land-use area shares. Temporal attributes reveal variations in activity durations and timings. For instance, cluster z_b = 11 represents infrequent, short-duration evening activities like visiting cafes or cinemas, while cluster z_b = 2 represents infrequent, long-duration morning activities like visiting park. Clusters z_b = 0, 9, and 10 involve longer-duration activities closer to the end of the day, such as picnics or exhibition visits.

Clusters z_b = 3 and 6 (See Fig. 8) indicate health and other services activities, occurring in stations with above-average land-use area shares for health and other services. Cluster z_b = 6 represents infrequent, short-lasting morning activities, possibly linked to routine medical checkups, while cluster z_b = 3 represents infrequent, long-lasting activities occurring at various times of the day, potentially associated with hospital visits or extended medical services.

Clusters z_b = 1 and 4 (See Fig. 10) could not be attributed to specific activity types based on their attribute characteristics. They exhibit broad temporal attribute distributions or lack a focus on a specific land-use category. These clusters may represent common activities across multiple types or minor activities that are challenging to infer from available information. They constitute a small proportion of all activities (14 % of all non-home & work activity type).

In this section, we presented the results of our clustering analysis for activity patterns in the Tehran subway system. We were able to identify several distinct patterns of activities based on the attributes of the stations where the activities occurred. We found that activities related to education, commercial activities, and recreational activities were among the most common patterns identified. Furthermore, we discovered two clusters that could be interpreted as health and other services activities. While we encountered some patterns that could not be classified into specific activity types, our results provide a useful framework for understanding the diverse range of activities that occur in subway system. While this process is fully unsupervised and was not cross validated with household surveys, it can still distinguish various activity patterns in terms of duration, start time, frequency, and land uses around the activity location. While we cannot guarantee the full accuracy of all activities categorized under each of the 14 defined activity patterns, clearly the historical data have given us distinguishable and interpretable clusters that can shed light on types of activities in the absence of a comprehensive manual travel surveys.

4.4. Impact of COVID-19 on activity patterns

In this section, we conduct an in-depth analysis of transit passengers' behavioral changes before and during the COVID-19 pandemic in Tehran using the model developed. The availability of data during the pandemic allowed us to compare the activity patterns of the same 10,000 individuals in January 2020 (pre-pandemic) and April 2020 (during the pandemic), providing valuable insights into the impact of the pandemic on travel behavior. Initially, we focused on the three major activity groups of home, work, and others, similar to our analysis before the pandemic.

Fig. 5 presents a comparison of attribute distributions among these clusters before and during the pandemic, shedding light on the changes in temporal patterns and activity frequencies. We can observe that the temporal patterns of work, home, and other activities for the same set of individuals remained relatively stable during the pandemic in terms of start time and duration. This indicates that the core temporal characteristics of these activities were largely unaffected by the pandemic. However, a significant proportion of individuals (45 %) eliminated their work and home activities, especially those infrequent activities. As such, the elimination of those infrequent activities causes an increase in average frequency of work activities done using subway system. This observation is aligned with the fact that less frequent activities were more significantly affected by the pandemic, indicating that discretionary and non-essential travel experienced a greater impact. To delve deeper into the impact of the pandemic on the other activity cluster, we

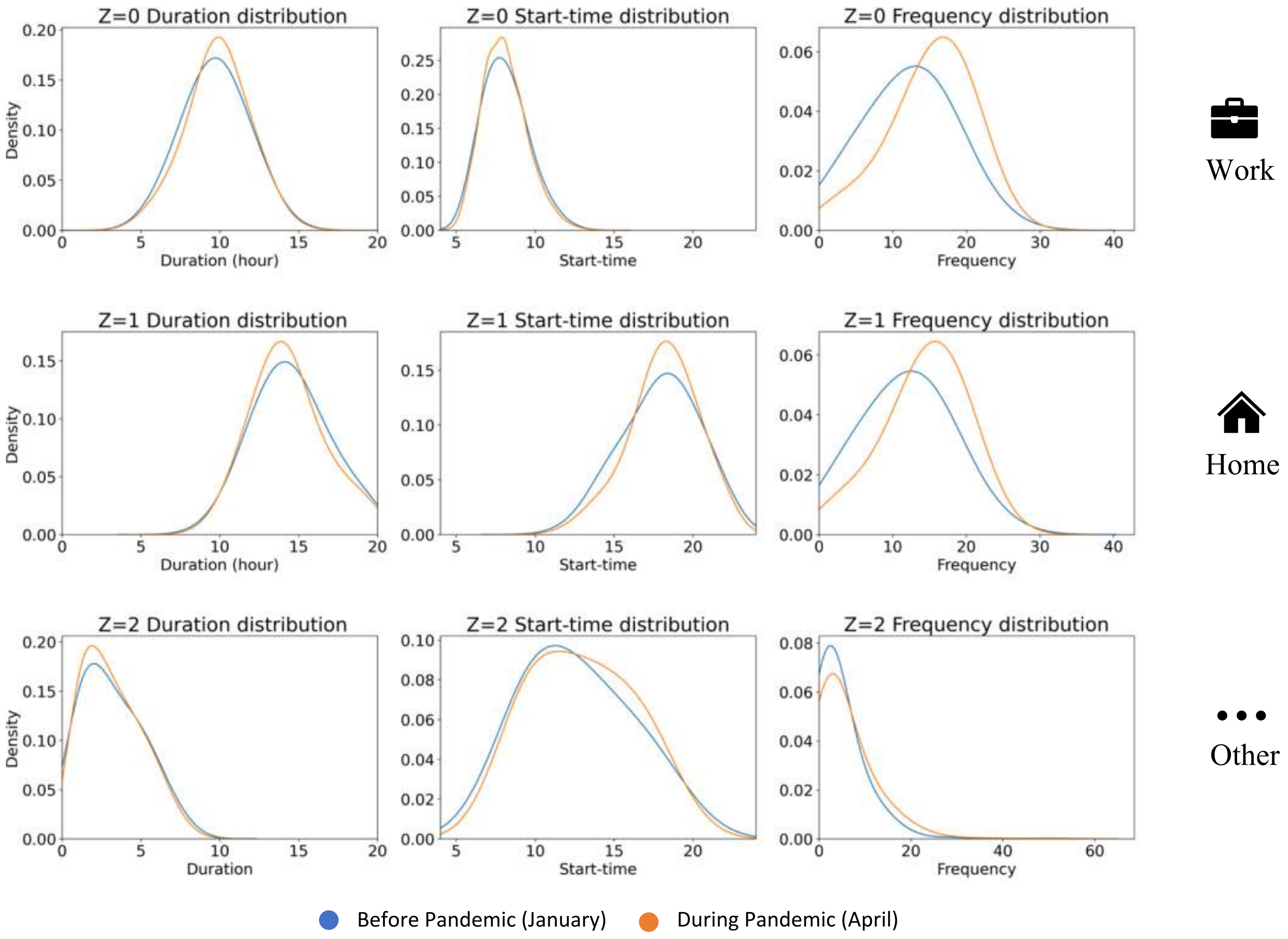

Fig. 5. Attributes distributions comparison for the 3 major clusters before and during the pandemic.

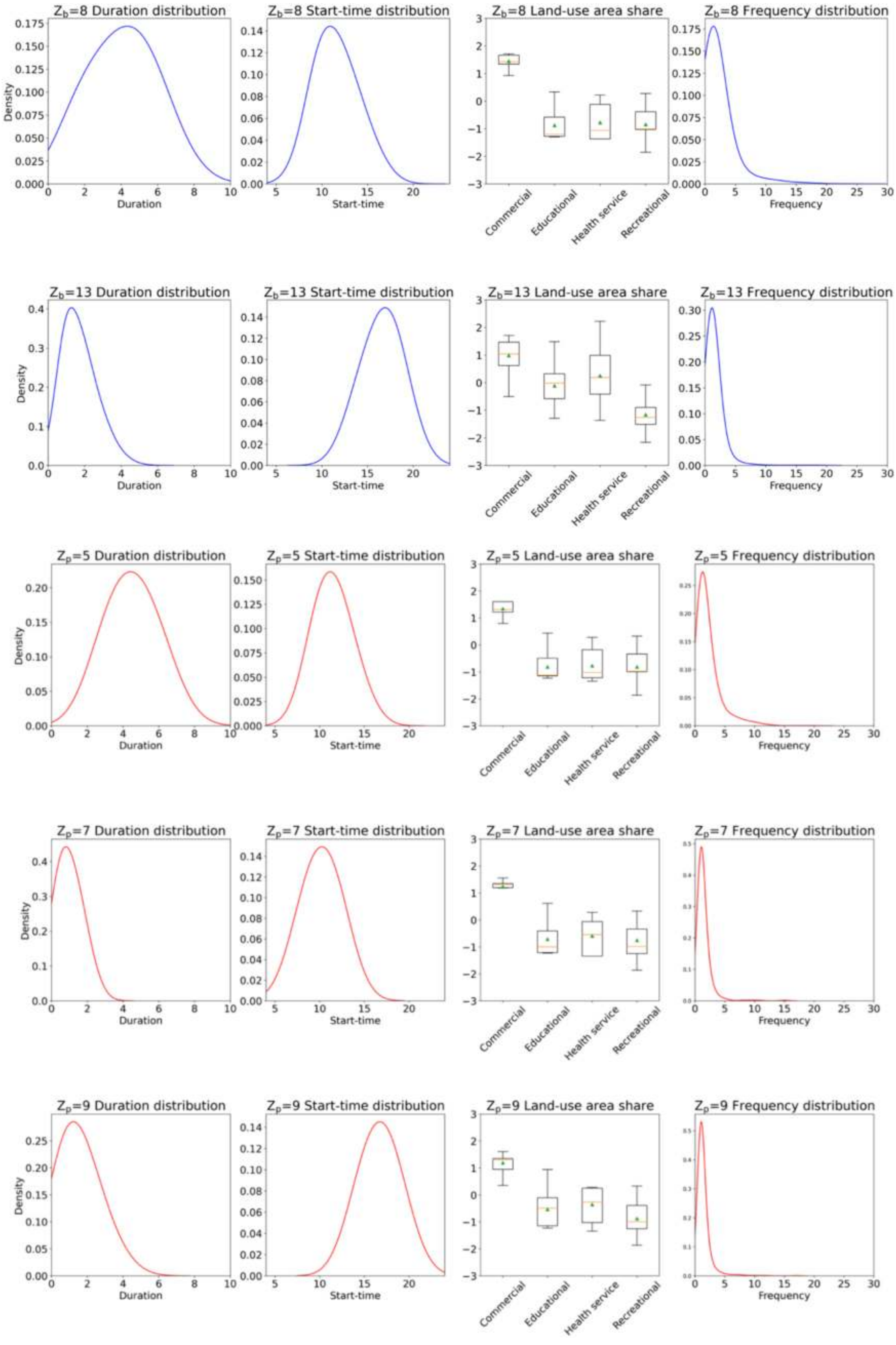

Fig. 6. Commercial activity patterns before (Z_b) and during pandemic (Z_p).

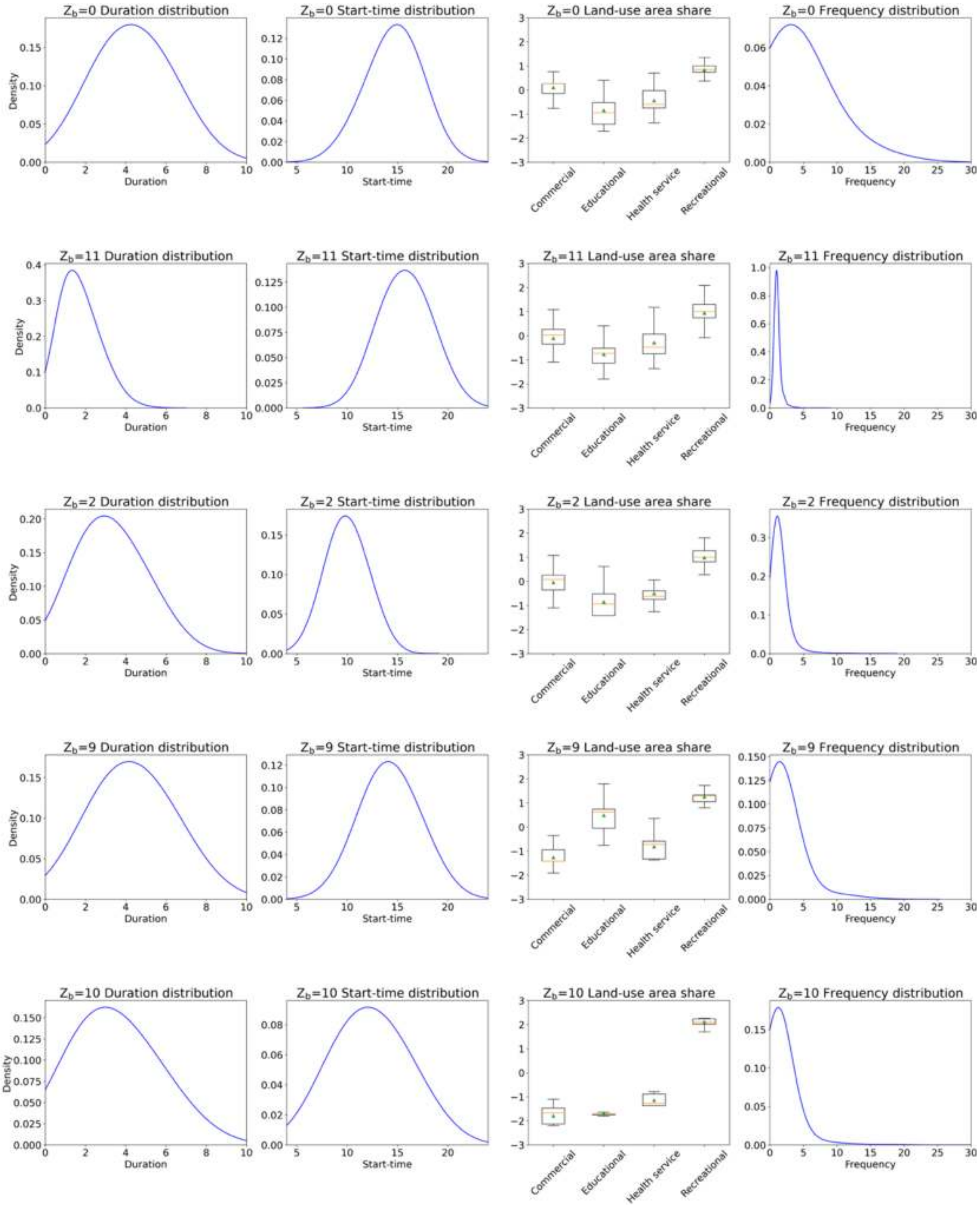

Fig. 7. Recreational activity patterns before (Z_b) and during pandemic (Z_p).

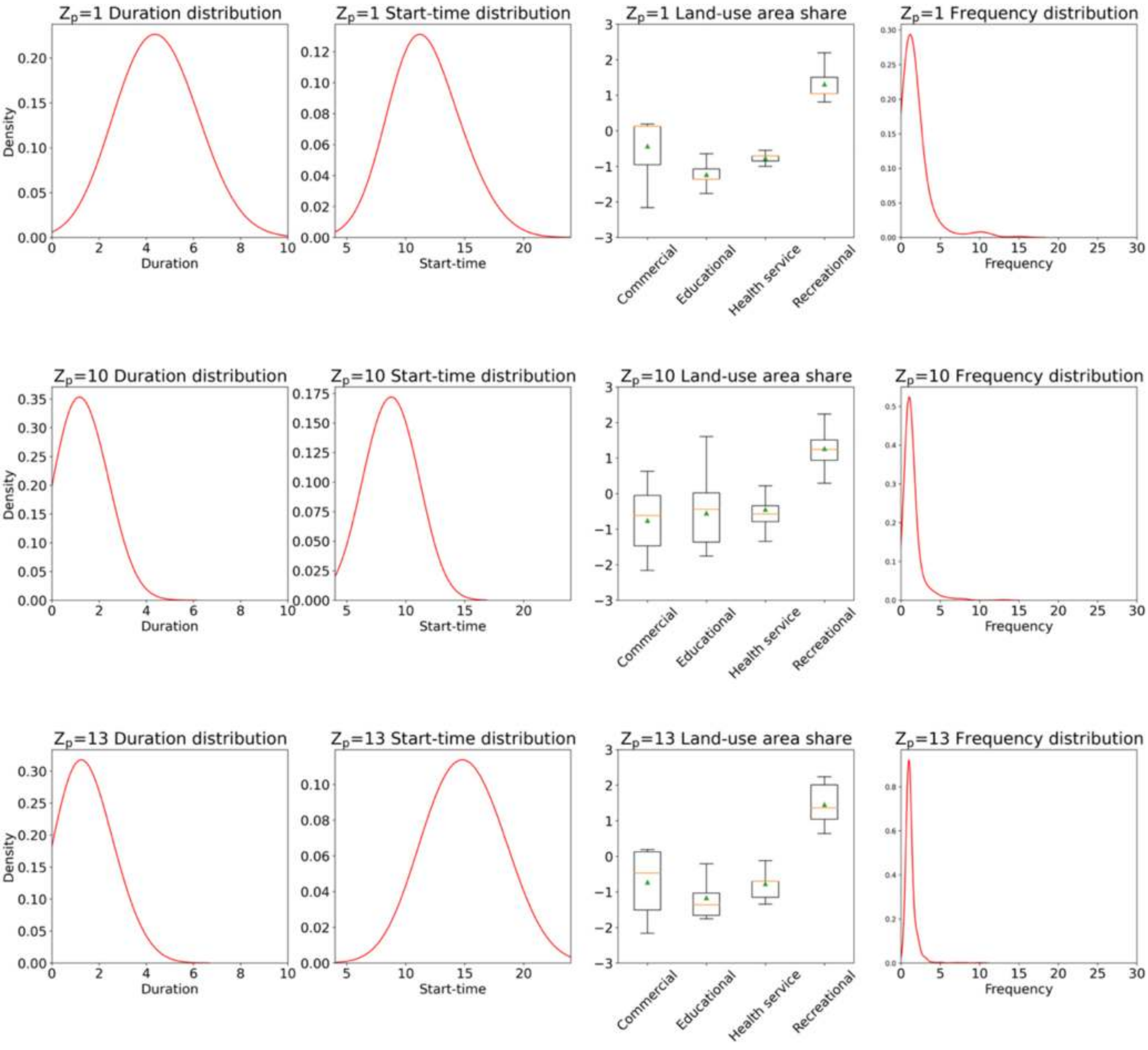

Fig. 7. (*continued*).

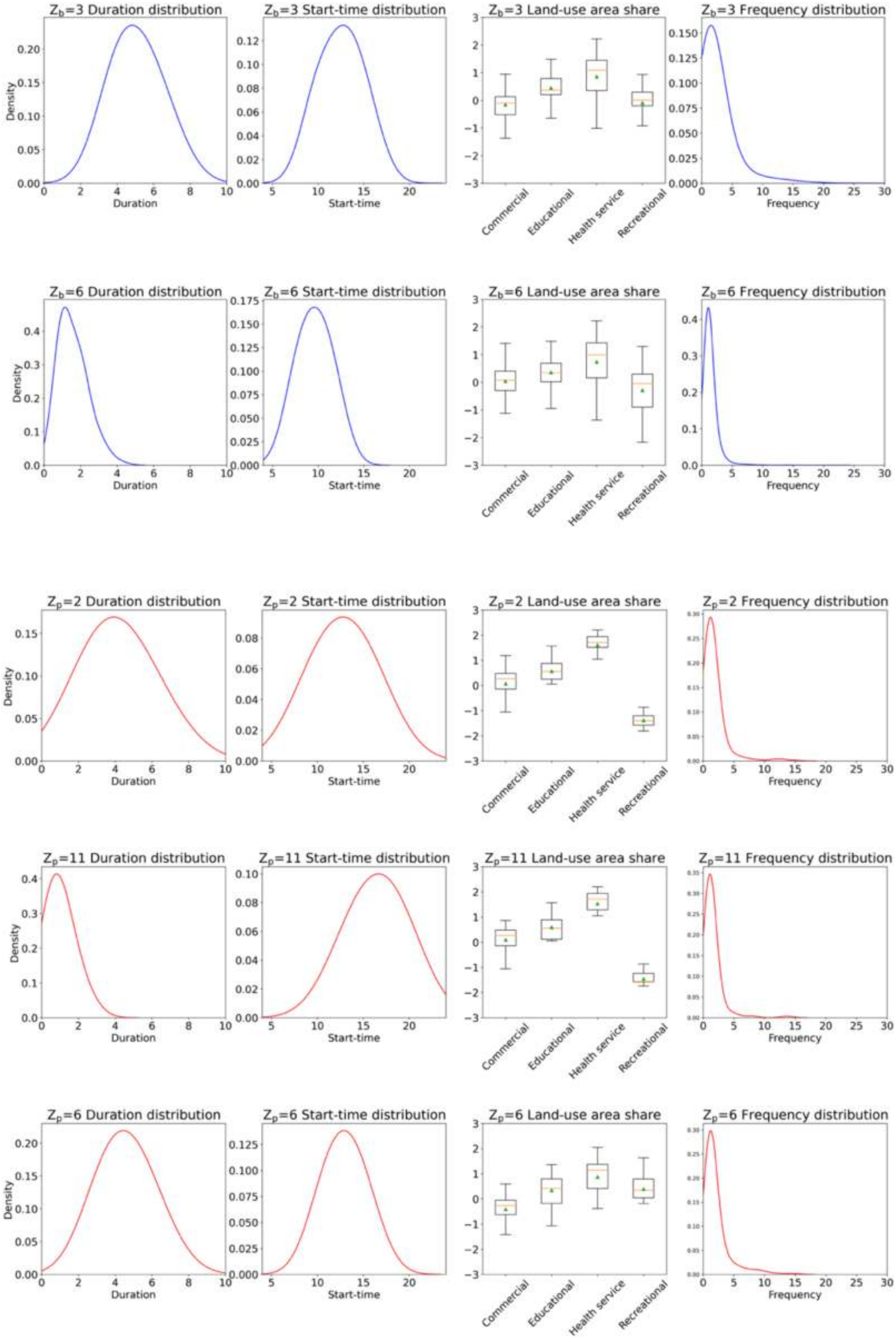

Fig. 8. Health and Other related activity patterns before (Z_b) and during pandemic (Z_p).

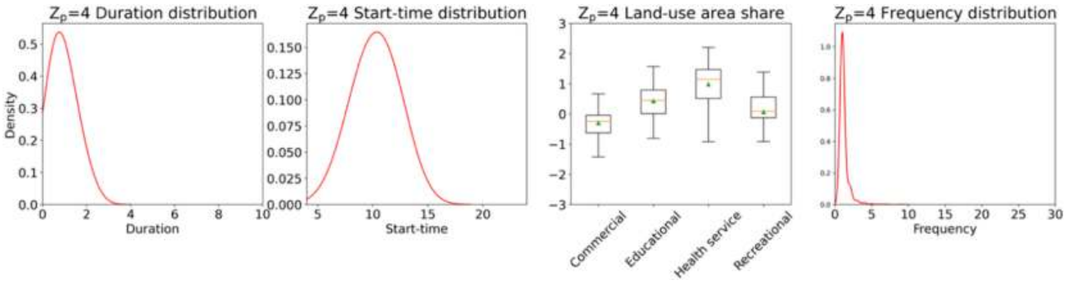

Fig. 8. (*continued*).

ran the LDA algorithm on the non-home & work activity cluster, as discussed in Section 4.3, to uncover more detailed activity patterns and infer activity categories based on attribute distributions. We then compared the patterns before and during the pandemic within each activity category to identify any significant changes. By examining these patterns, we can gain insights into the effects of the pandemic on different types of activities. The rest of this section includes discussion on explored the pattern changes in greater detail across all activity types. Subsequently, we explored whether the observed pattern changes align with the expected pandemic's impact on each activity type.

For commercial activities, the algorithm identified two major patterns before the pandemic, long-duration morning activities and short-duration evening activities. During the pandemic, a new pattern emerged with short-duration morning activities (See Fig. 6). Although the total number of commercial activities decreased by 60 %, the share of commercial activities among all non-home and work activity clusters remained close to pre-pandemic levels at around 28 %. This reduction in commercial activities can be attributed to the preference for online shopping and the desire to limit exposure in public places. However, there are still individuals who engage in in-person commercial activities, albeit with a slightly decreased duration, reflecting their efforts to minimize public interaction.

Recreational activities shown in Fig. 7, accounted for about 25 % of all non-home and work activities. These activities experienced a significant decrease during the pandemic, with the total number of such activities declining by approximately 76 %. The share of recreational activities among all non-home and work activities dropped to 17 % during the pandemic. This decline can be attributed to the limitations on public gatherings and the closure of recreational facilities. Individuals sought to expedite their activities to minimize exposure to COVID-19, as can be seen, where the average duration of long morning activities decreased from 4 h to 2 h (cluster 10 during pandemic).

In contrast, the patterns for health and other services activities remained relatively stable during the pandemic, reflecting their essential nature (See Fig. 8). However, a new pattern emerged during the pandemic for short-duration health and other services activities in the evening, possibly due to additional trips for COVID-19 testing and medication purchases. The number of activities in clusters 3 and 6 (during pandemic Z_p as shown in Fig. 8), which likely involve visits to health centers, decreased by 80 %. This may be due to concerns about infection transmission and advice for infected individuals to avoid using public transportation.

Education-related activities, particularly cluster 5 (during pandemic Z_p) in Fig. 9 representing long-duration morning activities, were significantly impacted by the shift to online learning. However, some activities still remained as observed in activity patterns in cluster 3 during the pandemic. These activities may be related to researchers and students who required access to university laboratories and also teachers who still required the facilities for remote teaching. Cluster 3 (during pandemic Z_p) exhibited a broader distribution of start times and durations, reflecting the flexibility for this type of activities which aligned with the nature of activities such as researchers and academicians compared to students who have fixed-time classes. Cluster 0 (during pandemic Z_p) represented short-duration activities distributed across a wide time range, likely associated with services near universities such as bookstores and educational offices.

Regarding changes in other activity types, notable shifts are evident. Start times for these activities have become more concentrated. Additionally, a comparison of the frequency distribution of discovered patterns before and during the pandemic (Fig. 10) indicates a decrease in the frequency of such activities. These observations align with the trend of people reducing non-essential activities during the pandemic period.

As mentioned earlier, the observations we derived from our output closely align with the impacts of the pandemic. This shows that our approach has the capacity to be applied to effectively detect activity pattern changes. Similar studies can be done for other short-term and long-term disruptions.

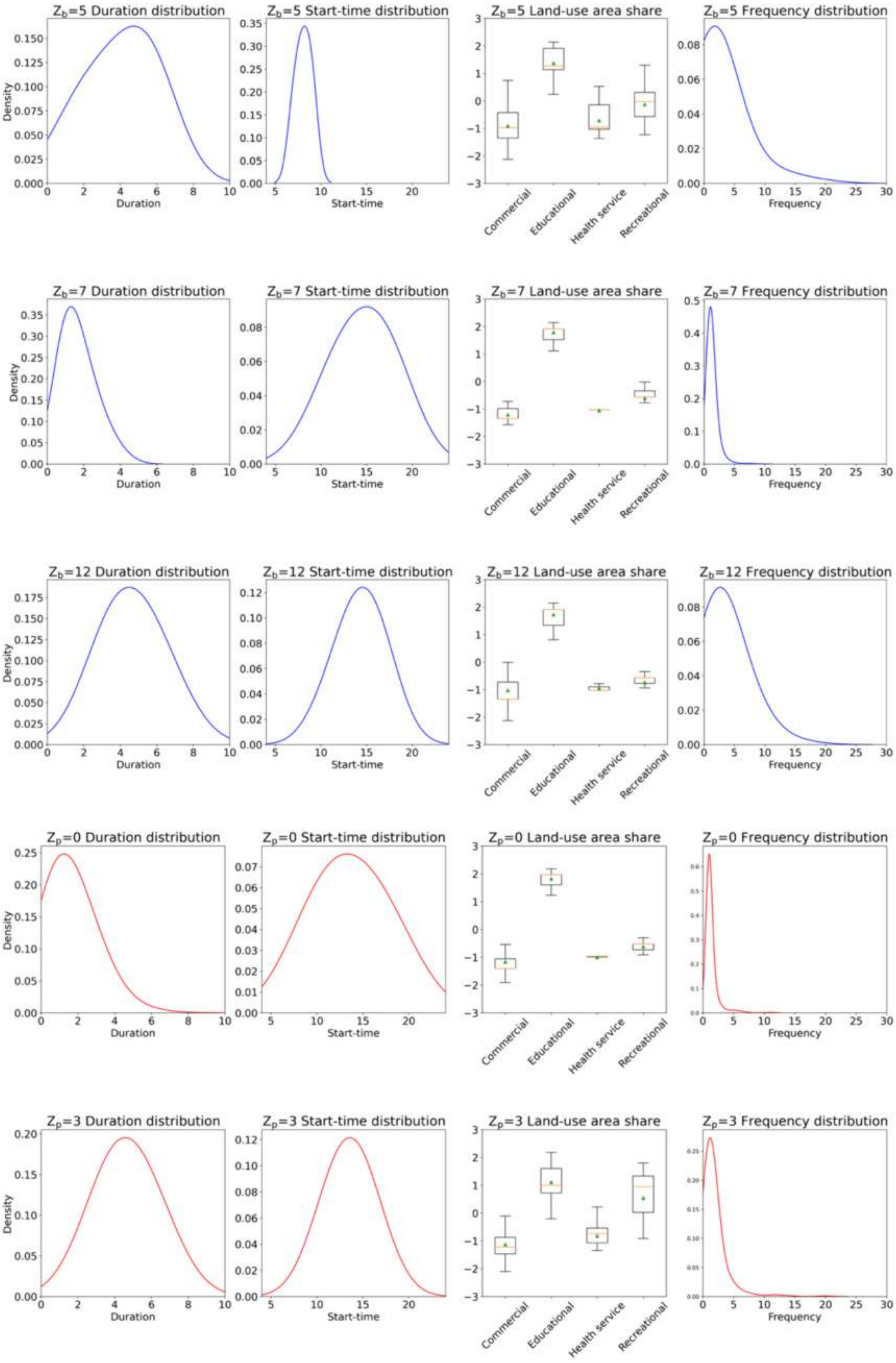

Fig. 9. Educational activity patterns before (Z_b) and during pandemic (Z_p).

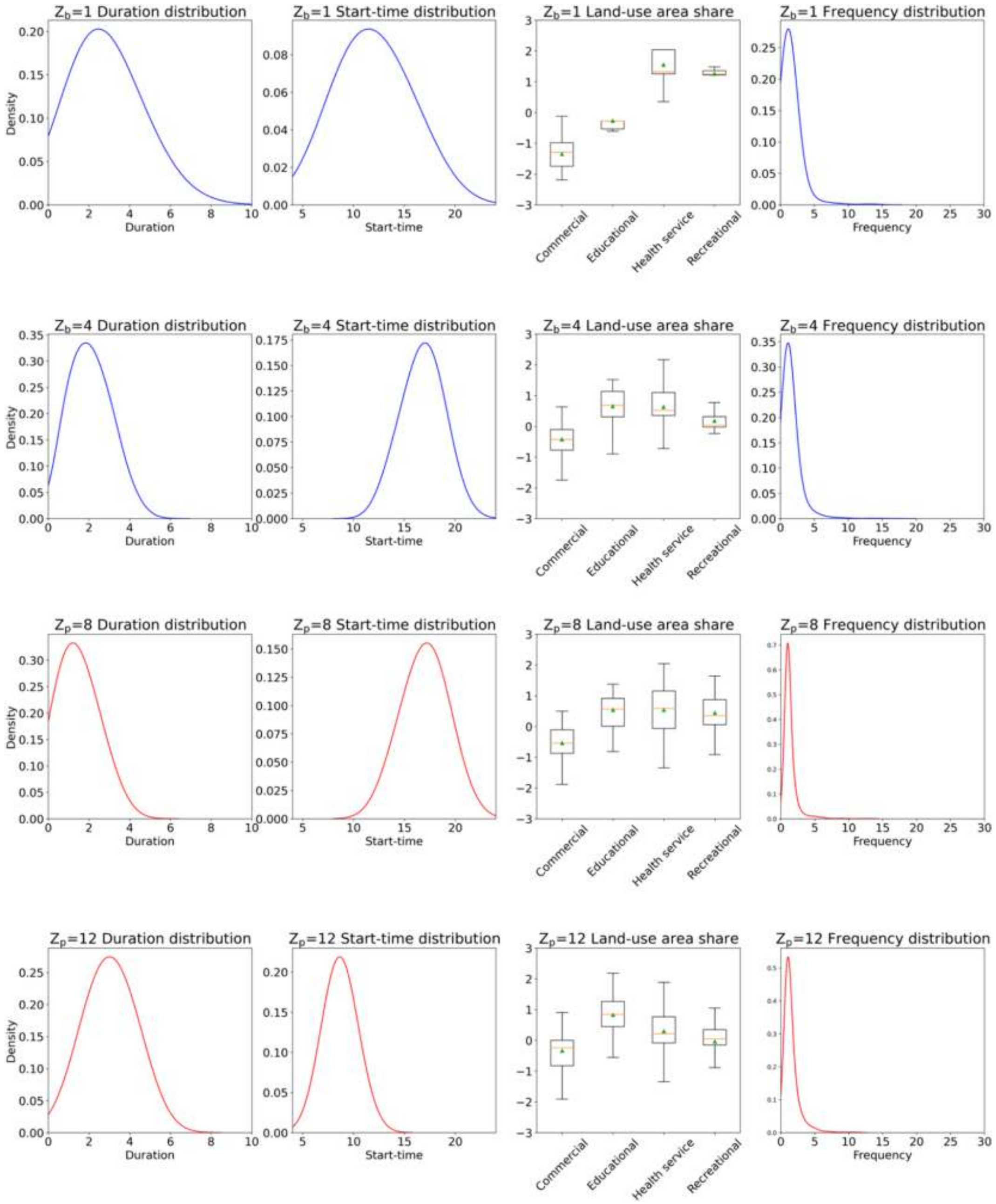

Fig. 10. Other activity patterns before (Z_b) and during pandemic (Z_p).

5. Conclusion

This study introduced a model to discover latent activities from passenger mobility by extending the LDA topic model using transit smart card and land-use data. By incorporating multiple dimensions of individual mobility, including location, start time, start day, and duration, we inferred hidden activity structures. The model differentiated between home and work-related activities and further classified activities into commercial, recreational, educational, and health service-related categories. During the COVID-19 pandemic, we analyzed the impact on these activity patterns and observed significant changes in the number of trips, start time, duration, and frequency for various activity clusters. The findings confirmed that activity pattern changes can be accurately captured consistent with known impact of COVID-19 pandemic on travel and activity patterns. This research contributes to our understanding of the spatiotemporal characteristics of transit passengers activities and provides valuable insights for urban planning and transportation policies. The proposed model offers advantages over traditional activity-based surveys, providing a cost-effective and scalable approach to studying travel behavior. It enables the characterization of individual activity patterns and facilitates user similarity measurement and cluster analysis. Additionally, the model can be used for mobility prediction, enhancing transportation planning and policy development. Implementing such model is extremely useful to detect pattern changes and also study the impact of different disruptions in travel and activity patterns.

This study, while pioneering in its approach to understanding urban activity patterns through the integration of transit smart card and land-use data, encounters several limitations. Firstly, a notable limitation revolves around the validation of our model results. Given that our methodology employs unsupervised learning techniques, there’s an inherent challenge in directly validating the inferred patterns of mobility without a standard benchmark or ground truth. Access to comprehensive travel survey data could significantly enhance the validation process. Unfortunately, the lack of access to such survey data in this study presents a limitation in validating the inferred activities with ground truth. However, the consistency of the model results in detecting activity pattern changes during the COVID-19 pandemic indicates that the model was able to capture significant shifts known to us. This consistency could serve as evidence of demonstrating model reliability and robustness in the absence of ground truth data. Another limitation concerns the computational cost associated with analyzing extensive datasets over extended periods and with a larger number of passengers. The computational resources required for processing and analyzing data at this scale are substantial. As the study seeks to expand its temporal scope or to include a greater number of passenger records with recurring analysis over time, the computational demands will proportionally increase, potentially limiting the feasibility of such expansions if computational resources are not adequate at the agency. Furthermore, while the study makes significant improvements in incorporating land-use information to understand its influence on activity patterns, there's room for enhancement in this domain. Specifically, a more refined approach to consider the varying attraction rates of different land-use categories could provide deeper insights. Accounting for these varying attraction rates in the analysis could offer a better understanding of how different land uses influence travel behavior and activity patterns. Future research can explore the application of the model to different datasets, incorporating additional dimensions and external factors to improve accuracy and understand the complex relationships between passenger behavior and the built environment. Further investigation into the impact of post-pandemic on urban mobility and activity patterns is also recommended. The model holds promise in its potential applications to mobility prediction methods, enabling the forecasting of an individual's forthcoming travel attributes and providing insights to guide transportation planning and policy formulation. Furthermore, the model can extend its utility to analyze the effects of various disruptions to mobility, such as the influence of escalating fuel prices, remote working, demographic changes in different region across diverse activity categories. This study contributes to the advancement of activity modeling and provides a foundation for future research in understanding travel behavior, predicting mobility patterns, and effective transportation policy interventions.

Funding sources

This work was financially supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant on "Urban Transit System Diagnosis, Monitoring, and Management using Mobility Sensing Data", Department of Civil Engineering and Schulich School of Engineering at the University of Calgary.

CRediT authorship contribution statement

Nima Aminpour: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Saeid Saidi:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

During the preparation of this work the authors used large language model ChatGPT in order to improve language and readability only. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Agard, B., 2009. Mining smart card data from an urban transit network. Encyclopedia of Data Warehousing and Mining, second ed. https://doi.org/10.4018/9781605660103.ch201.

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C: Emerg. Technol. 58, 240–250. https://doi.org/10.1016/j.trc.2015.02.018.

Allahviranloo, M., Recker, W., 2013. Daily activity pattern recognition by using support vector machines with multiple classes. Transp. Res. B: Methodol. 58, 16–43. https://doi.org/10.1016/j.trb.2013.09.008.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., Hickman, M., 2018. Public transport trip purpose inference using smart card fare data. Transp. Res. Part C: Emerg. Technol. 87 (January), 123–137. https://doi.org/10.1016/j.trc.2017.12.016.

Aminpour, N., 2024. Exploring Travel Behavior and Activity Patterns using Urban Transit Mobility Sensing Data. Graduate Studies at University of Calgary.

Anda, C., Erath, A., Fourie, P.J., 2017. Transport modelling in the age of big data. Int. J. Urban Sci. 21, 19–42. https://doi.org/10.1080/12265934.2017.1281150.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (3), 139–159. https://doi.org/10.1016/B978-0-12-411519-4.00006-9.

Chakirov, A., Erath, A., 2012. Activity identification and primary location modelling based on smart card payment data for public transport. In: 13th International Conference on Travel Behaviour Research (IATBR), July.

Chen, C., Gong, H., Lawson, C., Bialostozky, E., 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. Transp. Res. A Policy Pract. 44 (10), 830–840. https://doi.org/10.1016/j.tra.2010.08.004.

Chen, Z., Zhang, Y., Wu, C., Ran, B., 2019. Understanding individualization driving states via latent Dirichlet allocation model. IEEE Intell. Transp. Syst. Mag. 11 (2), 41–53. https://doi.org/10.1109/MITS.2019.2903525.

Chu, K.K.A., Chapleau, R., 2010. Augmenting transit trip characterization and travel behavior comprehension: multiday location-stamped smart card transactions. Transp. Res. Rec. 2183, 29–40. https://doi.org/10.3141/2183-04.

Come, E., Randriamanamihaga, N.A., Oukhellou, L., Come, E., Randriamanamihaga, N. A., Oukhellou, L., Spatio-temporal, P.A., 2014. Spatio-temporal analysis of dynamic origin-destination data using latent dirichlet allocation: application to Vélib ' bike sharing system of paris to cite this version: HAL Id: hal-01052951 spatio-temporal analysis of dynamic origin-destination data us. In: TRB 93rd Annual Meeting, November 2015, 19.

Devillaine, F., Munizaga, M., Trépanier, M., 2012. Detection of activities of public transport users by analyzing smart card data. Transp. Res. Rec. 2276, 48–55. https://doi.org/10.3141/2276-06.

Eagle, N., Pentland, A.S., 2009. Eigenbehaviors: Identifying structure in routine. Behav. Ecol. Sociobiol. 63 (7), 1057–1066. https://doi.org/10.1007/s00265-009-0739-0.

Faroqi, H., Mesbah, M., 2021. Inferring trip purpose by clustering sequences of smart card records. Transp. Res. Part C: Emerg. Technol. 127 (March), 103131 https://doi.org/10.1016/j.trc.2021.103131.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. 101(SUPPL. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

Han, G., Sohn, K., 2016. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. Transp. Res. B Methodol. 83, 121–135. https://doi.org/10.1016/j.trb.2015.11.015.

Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C: Emerg. Technol. 44, 363–381. https://doi.org/10.1016/j.trc.2014.04.003.

Kim, E.J., Kim, Y., Kim, D.K., 2021. Interpretable machine-learning models for estimating trip purpose in smart card data. Proc. Inst. Civil Eng.: Municipal Eng. 174 (2), 108–117. https://doi.org/10.1680/jmuen.20.00003.

Kuhlman, W., 2015. The construction of purpose-specific OD matrices using public transport smart card data. https://repository.tudelft.nl/islandora/object/uuid%3A7190712e-0913-4849-89ae-d1a1a88e66d2.

Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: a data fusion approach. Transp. Res. Part C: Emerg. Technol. 46, 179–191. https://doi.org/10.1016/j.trc.2014.05.012.

Lee, S.G., Hickman, M., 2014. Trip purpose inference using automated fare collection data. Public Transport 6 (1–2), 1–20. https://doi.org/10.1007/s12469-013-0077-5.

Li, Z., Xiong, G., Wei, Z., Zhang, Y., Zheng, M., Liu, X., Tarkoma, S., Huang, M., Lv, Y., Wu, C., 2021. Trip purposes mining from mobile signaling data. IEEE Trans. Intell. Transp. Syst. 1–13 https://doi.org/10.1109/TITS.2021.3121551.

Lin, Y., Xu, Y., Zhao, Z., Park, S., Su, S., Ren, M., 2023. Understanding changing public transit travel patterns of urban visitors during COVID-19: a multi-stage study. Travel Behav. Soc. 32 (April), 100587 https://doi.org/10.1016/j.tbs.2023.100587.

Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W., 2016. An overview of topic modeling and its current applications in bioinformatics. Springerplus 5 (1). https://doi.org/10.1186/s40064-016-3252-8.

Ma, X., Wu, Y.J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C: Emerg. Technol. 36, 1–12. https://doi.org/10.1016/j.trc.2013.07.010.

Mo, B., Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2022. Individual mobility prediction in mass transit systems using smart card data: an interpretable activity-based hidden

Markov approach. IEEE Trans. Intell. Transp. Syst. 23 (8), 12014–12026. https://doi.org/10.1109/TITS.2021.3109428.

Montini, L., Rieser-Schüssler, N., Horni, A., Axhausen, K., 2014. Trip purpose identification from GPS tracks. Transp. Res. Rec. 2405, 16–23. https://doi.org/10.3141/2405-03.

Nassir, N., Hickman, M., Ma, Z.L., 2015. Activity detection and transfer identification for public transit fare card data. Transportation 42, 683–705.

Peng, C., Jin, X., Wong, K.-C., Shi, M., Lio, P., 2012. Lasp-1 regulates podosome function. PLoS One 7 (4), 1–10. https://doi.org/10.1371/Citation.

Rasiwasia, N., Vasconcelos, N., 2013. Latent Dirichlet allocation models for image classification. IEEE Trans. Pattern Anal. Mach. Intell. 35 (11), 278–279.

Sari Aslam, N., Cheng, T., Cheshire, J., 2019. A high-precision heuristic model to detect home and work locations from smart card data. Geo-Spatial Inf. Sci. 22 (1), 1–11. https://doi.org/10.1080/10095020.2018.1545884.

Sari Aslam, N., Ibrahim, M.R., Cheng, T., Chen, H., Zhang, Y., 2021. ActivityNET: neural networks to predict public transport trip purposes from individual smart card data and POIs. Geo-Spatial Inf. Sci. 24 (4), 711–721. https://doi.org/10.1080/10095020.2021.1985943.

Sato, Y., Maruyama, T., 2020. Modeling the rounding of departure times in travel surveys: comparing the effect of trip purposes and travel modes. Transp. Res. Rec. 2674 (10), 628–637. https://doi.org/10.1177/0361198120935435.

Wang, J., Chen, X., Chen, Z., Mao, L., 2016. Cluster algorithm based on LDA model for public transport passengers' trip purpose identification in specific area. In: 2016 IEEE International Conference on Intelligent Transportation Engineering, ICITE 2016. pp. 186–192. https://doi.org/10.1109/ICITE.2016.7581331.

Xiao, G., Juan, Z., Zhang, C., 2016. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transp. Res. Part C: Emerg. Technol. 71, 447–463. https://doi.org/10.1016/j.trc.2016.08.008.

Yu, C., Li, H., Xu, X., Liu, J., Miao, J., Wang, Y., Sun, Q., 2021. Data-Driven approach for passenger mobility pattern recognition using spatiotemporal embedding. J. Adv. Transp. 2021, 13–20. https://doi.org/10.1155/2021/5574093.

Zhang, Y., Sari Aslam, N., Lai, J., Cheng, T., 2020. You are how you travel: a multi-task learning framework for Geodemographic inference using transit smart card data. Comput. Environ. Urban Syst. 83 (June), 101517 https://doi.org/10.1016/j.compenvurbsys.2020.101517.

Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2020. Discovering latent activity patterns from transit smart card data: a spatiotemporal topic model. Transp. Res. Part C: Emerg. Technol. 116(July 2019), 102627. https://doi.org/10.1016/j.trc.2020.102627.

Zhao, X., Li, Z., Zhang, Y., Lv, Y., 2020. Discover trip purposes from cellular network data with topic modelling. IEEE Intell. Transp. Syst. Mag. September 2020. https://doi.org/10.1109/MITS.2020.3014111.

Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2020b. Discovering latent activity patterns from transit smart card data: a spatiotemporal topic model. Transp. Res. Part C: Emerg. Technol. 116 (February), 102627 https://doi.org/10.1016/j.trc.2020.102627.

Zou, Q., Yao, X., Zhao, P., Wei, H., Ren, H., 2018. Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. Transportation 45 (3), 919–944. https://doi.org/10.1007/s11116-016-9756-9.