Maeda *et al. J Big Data* (2019) 6:4
https://doi.org/10.1186/s40537-019-0168-5

Journal of Big Data

RESEARCH Open Access

Detecting and understanding urban changes through decomposing the numbers of visitors' arrivals using human mobility data

Takashi Nicholas Maeda[1*], Narushige Shiode[1,2], Chen Zhong[2], Junichiro Mori[1] and Tetsuo Sakimoto[1]

*Correspondence:
maeda@ipr-ctr.t.u-tokyo.ac.jp
[1] The University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo 113-8656, Japan
Full list of author information is available at the end of the article

Abstract

In recent years, mobility data from smart cards, mobile phones and sensors have become increasingly available. However, they often lack some of the key information including the purposes of trips for each individual user. Information on trip purposes is crucial for projecting the future travel patterns as well as understanding the characteristics of each area of a city and how it is changing. This paper proposes a method called EAT-CD (Extraction of Activity Types and Change Detection). It estimates the volume of passengers by activity types (e.g. commuting, leisure) using non-negative matrix factorization and detects changes in the number of visitors for each activity (e.g. increase in shopping trips triggered by the development of a new commercial facility). Validity of EAT-CD is tested through empirical analysis using smart card data of public transportation in Western Japan. The results showed that EAT-CD is effective in deriving activity patterns, which showed strong correlation with travel survey data. The results also confirmed that EAT-CD detects changes in travel patterns (e.g. start and end of semesters) and land uses (e.g. establishment of new facilities).

Keywords: Change detection, Human mobility, Non-negative matrix factorization, Public transportation

Introduction

Trips within and between cities are manifested through the need to access places and participate in activities [1]. They range from the daily commute to workplaces to ad hoc excursions. Analyzing the patterns of such human mobility marks an important step towards understanding human activities in a city and, thereby, help planning and managing public transport and road traffic. To this end, urban planners, social scientists and real-estate developers have often relied on travel survey data (e.g. Axhausen et al. [2]). While survey data reflects human mobility and sheds light on the trends of human activities in a city, they are costly and time-consuming to collect.

In recent years, mobility data from smart cards, mobile phones and sensors have become increasingly available [3]. The process of collecting such data is largely automatic and requires minimum effort, yet they reflect real-time human movements within urban space. Given the extensive and comprehensive nature of mobility data, they are expected to help contribute to the planning and improvement of cities in

developed and the developing countries alike. Indeed, the World Bank [4] and the United Nations [5, 6] emphasise the scope to utilise such mobility data for covering population in developing countries. These include cases of the TfL (Transport for London) Oyster Card data being used for identifying deprived areas within London, UK [7], and mobile phone data in Rwanda being utilised for understanding patterns of local migration within the country [8].

However, mobility data often lack some of the key information, and these include the purposes of trips for each individual user. Information on trip purposes is crucial for estimating the number of visitors for each type of activity (e.g. commuting, leisure), and their changes over time would help identify the characteristics of each area and how it is changing.

Previous studies have estimated trip purposes, land uses, and activities using individual travel patterns [9, 10] and other secondary information such as household survey data and POI (point of interest) in the area [11, 12]. However, these datasets also tend to be unavailable or inaccurate. In order to estimate the trip patterns and monitor urban changes without such limitations, a new method is needed.

Changes in urban space such as the development of a new commercial facility often have different impact on human mobility by the trip purpose (e.g. commuting, leisure). This makes it crucial to understand the break-down of the number of human flows to understand urban changes. By the same token, decomposing human flow in a city would enable us to detect and understand such urban changes.

This paper investigates a method for estimating the numbers of visitors for different trip purposes through mobility data collected through passively monitoring the passenger flow that reflect the actual human movement (e.g. smart card data of public transportation). In this study, we use the smart card data of public transportation in Japan. Specifically, we propose a method called EAT-CD (Extraction of Activity Types and Change Detection) that estimates the volume of passengers by each activity and detects changes in the number of visitors for each activity; e.g. increase in shopping trips triggered by the development of a new commercial facility.

Our research questions are summarised as below.

- *RQ1* Can we develop a method that estimates the numbers of visitors to each place by their trip purposes using mobility data (e.g. smart cards records from public transportation) without individual trip records?
- *RQ2* Can we develop a method that detects changes in the activity trends from the estimates of the number of visitors by each activity?

The remainder of this paper comprises the following. "Related studies" section reviews previous studies, followed by the introduction of the proposed method ("Methods" section) for estimating the volume of trips by activities, and detecting changes in the trend of urban activities. "Dataset" section explains the dataset used and the analysis carried out in the empirical study. "Results and discussions" section discusses the results of the analysis, and "Conclusion" section brings the paper to a conclusion.

Related studies

There have been many studies on estimating trip purposes, land uses, and activities, and understanding urban changes, on the basis of human mobility data. The following sections give an overview of studies on inference of purpose, land use, and activity, with a focus on understanding urban changes, and a gap in the literature and how we might address it.

Inference on trip purposes, land uses, and activities

Trip purpose, land use, and activity are known to affect one another, which means that studies on their inference also have a considerable overlap, employing similar types of data and methodologies.

Inference based on individual travel patterns and additional secondary data

Individual travel patterns have been used by several studies for inferring the respective trip purpose [9, 10, 13–16]. For instance, Alexander et al. [9] infer trip purpose from call detail records (CDRs) which are collected through the use of mobile phones that contain time-stamped geo-coordinates. They estimate the location of each mobile phone user and classify them into their home, work, and other places depending on the frequency of observation, day of the week, and time of the day.

Others use POI data [17–20], as they tend to offer information on specific trip purpose, land use and activity. For instance, Wang et al. [18] infer subway station functions by applying the Doc2vec model [21] to smart card data and POI data. Zhong et al. [20] propose a method for inferring building functions in Singapore using smart card data and POI data.

Supplementary information such as land use data and household survey data are also used for inferring land use, activity, and trip purpose [11, 12, 22]. For example, Long et al. [22] combine smart card data with household travel surveys, as well as a parcel-level land use map to identify job-housing locations and commuting trip routes in Beijing.

The above studies use detailed individual trip records which are usually not open to the public. In addition, they use additional secondary data that tend to be inaccurate or not frequently updated. In this study, we develop a method that uses hourly number of human inflows in each area.

Inference based on patterns of temporal distribution of population or trips

Another strand of literature has focused on inferring the land uses and activities within each area using patterns of temporal distribution of population and trips. The advantage of using such methods is that they can identify daily land use of each area without any additional information such as POIs. For instance, Nishi et al. [23] extract area-by-area and daily land use patterns using location data obtained from mobile phones. Their method creates a 24-dimesional vector for each area and each day, which retains information about hourly numbers of human inflow. Each vector is then normalised by dividing all 24 elements by the total number of the elements. Their study uses the infinite Gaussian mixture model (GMM) which incorporates the Dirichlet process (DP) to cluster the vectors, which are labeled collectively with the

respective land use type. Similarly, Frias-Martinez et al. [24] use the number of posts of Twitter users for clustering the same land use types within each area. Each area is vectorised by 144 elements, namely the numbers of posts every 20 min on weekdays and weekends. Land use of each area is estimated by applying spectral clustering to the vectors. Chen et al. [25] propose to delineate areas with urban functions based on social media data aggregated to the building-block level.

The above studies identify land uses/activities without individual trip records or any additional information. However, they label single land use or activity to each area, and their methods cannot capture mixed land use. Therefore, our study develops a method for capturing multiple activities in each area.

Analyses of urban changes

Use of mobility data for interpreting changes in urban space and its usage is becoming an increasingly prominent research topic. Studies on the topic often apply analytical methods based on machine learning or spatial networks to measure changes in the landscape and usage of urban environment.

These studies are often confined by the limited range of attributes available in the mobility data. In order to estimate detailed patterns and changes in urban space, we need to find a way to decompose the human flows or population into different categories of activities and travel purposes such as commuting and leisure. Several studies have focused on this aspect. For instance, Fan et al. [26] propose a tensor factorization approach to modeling city dynamics. The study utilises non-negative tensor factorization (NTF) [27] to decompose a human flow tensor obtained from GPS log data into basic life pattern tensors such as commuting, working, and entertaining. It applies the method for modeling the fluctuation in human flow before and after the Great East Japan Earthquake. Another study (Wang et al. [28]) models time-evolving traffic networks into a 3-order origin-destination-time tensor, which detects the spatial clusters, temporal patterns and the associations among such networks.

Spatial network analysis is another important method for detecting the dynamics of urban structure. Zhong et al. [29, 30] propose quantitative measures to evaluate the centrality of locations. Their study applies this method to mobility data collected in Singapore and conclude that the city-state has been rapidly transforming to adopt a polycentric urban form.

In our study, we develop a method to detect changes in each type of activity (e.g. commuting and leisure) within each area of a city. For example, the number of visitors for shopping would increase in a place where a shopping center opens. Our study aims to detect such changes by estimating the break-down between such activities within each area.

Gap in the literature

Although various studies have attempted to estimate trip purposes and land uses as well as changes in human activities and urban landscape in general, to our knowledge, no study has proposed a method for measuring and detecting changes in the trend of activities within each area.

Our research was originally motivated by Nishi et al. [23] and Fan et al. [26]. In Nishi et al. [23], the daily land use of each area is derived from the number of

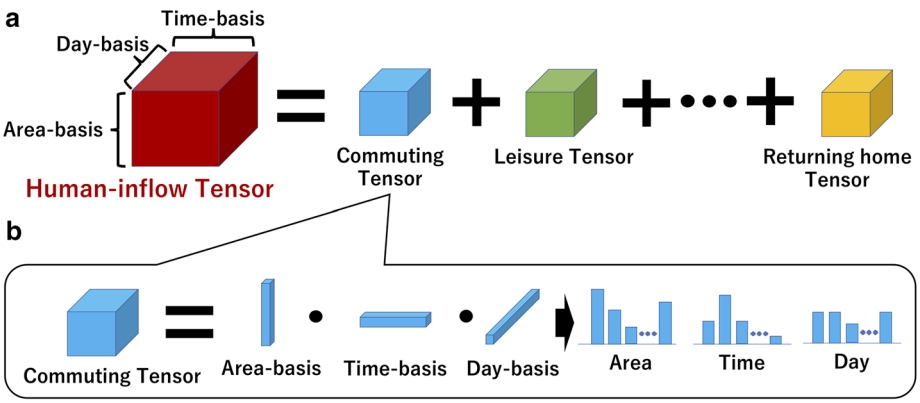


Fig. 1 Application of non-negative tensor factorization to human mobility data proposed in a previous study [26]

population recorded at each hour. However, most areas in a city may have mixed land use and activities, and their method would not be suitable for estimating the break-down between such activities within each area. On the other hand, Fan et al. [26] decompose human population into several activities by using non-negative tensor factorization (Fig. 1). Their tensor consists of area-basis, time-basis and day-basis, which respectively denote population by area, hour and day. This tensor is then converted and decomposed into a set of tensors. Each of the new tensors represents an activity type such as commuting and leisure (Fig. 1a). It allows us to estimate the temporal and spatial distribution of each activity (Fig. 1b), and have an overview of the trend of each activity across the entire study area. However, as their method estimates the temporal and spatial patterns of each type of activity at the aggregate level only, it is impossible to see the trend of each activity for each area separately. For instance, the case study discussed in their study features the impact of the Great East Japan Earthquake on human activity, but the method cannot show area-specific trends.

To understand the trend of each activity by area in a city, we need a new method that satisfies the following criteria: (1) it incorporates the temporal patterns of the distribution of the population within each area; (2) it decomposes the distribution of population into some activity types; and (3) it shows the chronological changes in activities for each area.

This paper aims to propose a method that estimates the activity trends within each area of a city by extending the method proposed by Fan et al. [26]. The method will also be designed to detect changes in each activity for each area. The method will be tested with an empirical case study that utilises smart card data of public transportation in western Japan, but the proposed method is applicable to any other mobility data including GPS log data.

Methods

In this section, we propose a method called EAT-CD (Extraction of Activity Types and Change Detection) for detecting and estimating changes in the trends of activities within each area of a city. Figure 2 shows the procedure involved in carrying out EAT-CD. First, EAT-CD decomposes the temporal distribution of the numbers of visitors'

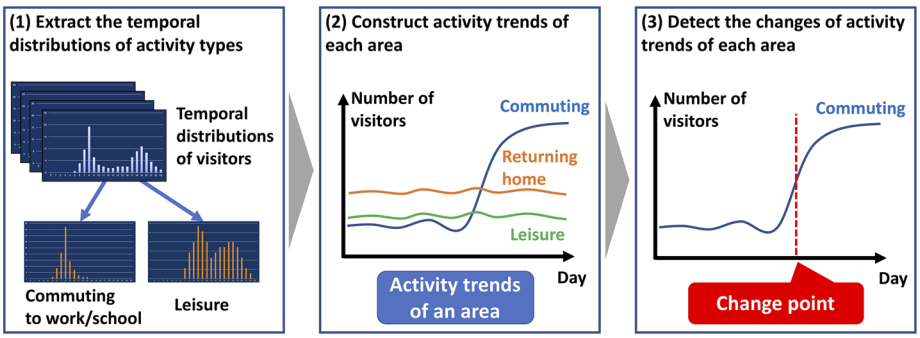


Fig. 2 A flow-chart of the processes involved in carrying out our proposed method

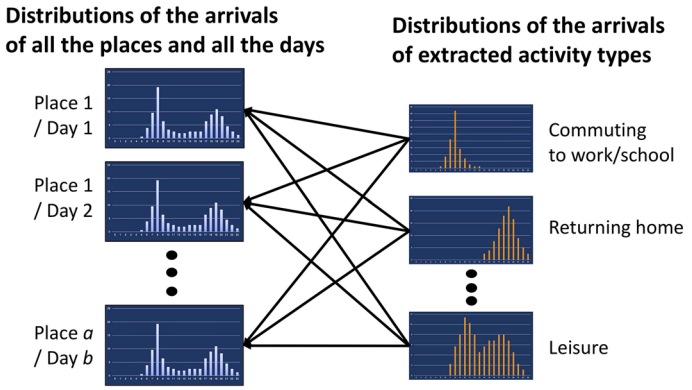


Fig. 3 Decomposition of the number of visitors' arrivals into different activity types

arrivals at each place for each day into a set of activity types such as commuting and leisure (Fig. 2-(1)). The underlying assumption is that such distribution is represented by the linear sum of these basic distributions. This allows us to construct the trend of each activity type in each place (Fig. 2-(2)). EAT-CD automatically detects changes in the trend of each activity type in each place. EAT-CD may detect changes caused by events such as the construction of a new commercial facility and the beginnings and the ends of school terms (Fig. 2-(3)).

Decomposing the temporal distribution of visitors' arrivals into activity types

We assume that the temporal distribution of the number of visitors' arrivals at each place for each day can be assumed to consist of superposition of multiple basic distributions, as shown in Fig. 3. For example, if the area containing a railway station is characterised as a mixture of a business area and a residential area, the number of passengers' arrivals on a weekday may have an acute concentration around 8 AM to 9 AM, and a more gradual increase around 6 PM. The distribution can be regarded as the superposition of two basic distributions, namely the passengers commuting/schooling to this area in the morning and the passengers returning to their home in this area in the evening. The temporal distribution of passengers' arrivals at every

place on every day can be expressed as the linear sum of a finite number of distributions that collectively represent the activity types such as commuting/schooling and leisure.

Figure 1 illustrates the framework of the method proposed by Fan et al. [26]. By extending their method, we propose a method detailed in Fig. 4. First, we create a matrix which consists of (area × day)-basis and time-basis, and each element of the matrix denotes the number of visitors to an area during one time-slice in each day. Each row denotes an aggregate of all visitors to one area in a day across all time-slices, and each column denotes visitors to all areas during one time-slice in a day. This matrix is then decomposed into a set of matrices. Each of the new matrices represents a unique activity type such as commuting or leisure (Fig. 4a). Each matrix is represented as a matrix product of a single-column vector and a single-row vector (Fig. 4b). The single-column vector provides information about the number of people who visit each area each day for a particular activity (Fig. 4c). In this manner, we obtain the number of visitors to each area for each activity. Finally, we obtain the break-down of the number of visitors' arrival (Fig. 4d). Our method uses non-negative matrix factorization (NMF) [31] for decomposing the number of visitors.

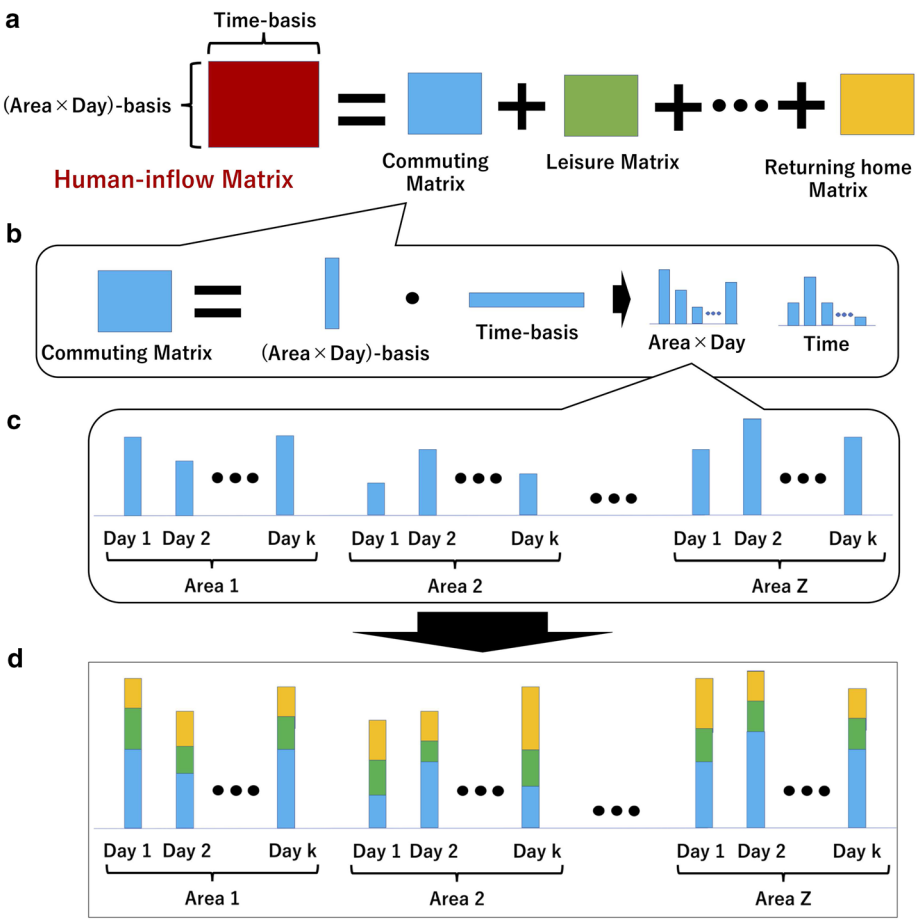


Fig. 4 Proposed application of non-negative matrix factorization to human mobility data

The objective of non-negative matrix factorization

The purpose of non-negative matrix factorization is to represent a matrix as the product of two non-negative matrices. The variables we use here are listed in Table 1. We create matrix $\mathbf{V}$ that has information about the number of visitors for each place, each day, and each time-slice. We discretise the time by dividing 1 day into n time-slices. If the length of one time-slice is 1 hour, then $n = 24$. Let a denote the number of all places in the data, b denote the number of days in the data, and define m as $m = ab$. Let $m \times n$ matrix $\mathbf{V}$ denote the numbers of visitors' arrivals at all a number of places for b number of days across n instances of time-slices (i.e. each row vector of $\mathbf{V}$ indicates a place and a day, and each column vector indicates a time-slice). By using non-negative matrix factorization, Matrix $\mathbf{V}$ is represented as the product of the two non-negative matrices, $\mathbf{W}$ and $\mathbf{H}$: $\mathbf{V} = \mathbf{WH}$, as shown in Fig. 5. Each row vector of matrix $\mathbf{H}$ provides a break-down of the distribution of the numbers of visitors' arrivals for a trip activity, and is expressed in a vector form whose sum of all its elements is 1. The number of columns of matrix $\mathbf{W}$ and the number of rows of matrix $\mathbf{H}$ are equal to the number of activity types, and this number is determined arbitrarily. In this research, the number is increased by one unit at a time. When two similar basic distributions are detected, the process of increasing

Table 1 Definition of variables

Variable	Definition
a	The number of places in the data (e.g. the number of stations in the smart card data of public transportation)
b	The number of days in the data
m	$a \times b$
n	The number of time-slices in 1 day. For example, if the length of one time-slice is 1 hour, $n = 24$. If the length of one time-slice is 30 minutes, $n = 48$
$\mathbf{V}$	A matrix that has information about the number of visitors for each day, each place, and time-slice. Each row indicates a specific day and place. Each column indicates a specific time-slice
$\mathbf{W}$	A matrix that has information about the inferred number of visitors for each place and each day by activity type. Each row indicates a specific day and station. Each column indicates a specific activity type
$\mathbf{H}$	A matrix that has information about time-series distribution of visitors for each activity type. Each row indicates a specific activity type. Each column indicates a specific time-slice

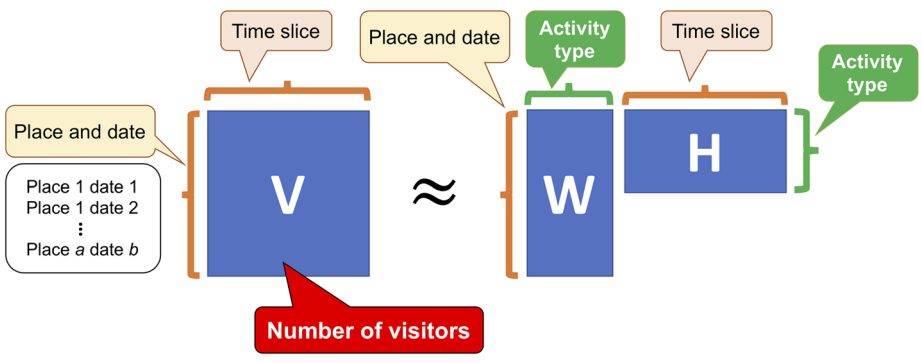


Fig. 5 A diagram showing the objective of NMF in this paper. Matrix $\mathbf{V}$ indicates the numbers of visitors' arrivals across all places among all time-slices. Matrix $\mathbf{V}$ is represented as the product of the two non-negative matrices, $\mathbf{W}$ and $\mathbf{H}$

the number is stopped. The number is decided as the maximum value of matrix **H** under the condition that no similar data on trip distributions are available.

The algorithm of non-negative matrix factorization

We use an algorithm of NMF originally proposed by Lee and Seung [31] to decompose the temporal distribution of the numbers of visitors' arrivals at each place each day. The following explanation of the algorithm of NMF is based on Sawada et al. [32]. The objective is to minimize the difference between matrix **V** and matrix **WH**. The difference is defined by Eq. 1.

$$D(\mathbf{V}, \mathbf{WH}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d(x_{ij}, \mathbf{t}_i^{\mathrm{T}} \mathbf{v}_j) \tag{1}$$

where x_{ij} denotes an element of **V**, $\mathbf{t}_i$ denotes the i-th row vector of **W**, and $\mathbf{v}_j$ denotes the j-th column vector of **H**.

There are three popular definitions of d, namely Euclidean distance [31], Kullback-Leibler distance [31], and Itakura-Saito distance [33]. Euclidean distance is used in our analysis for the sake of simplicity.

$$d_{\mathrm{Eu}}(x_{ij}, \mathbf{t}_i^{\mathrm{T}} \mathbf{v}_j) = (x_{ij} - \mathbf{t}_i^{\mathrm{T}} \mathbf{v}_j)^2 \tag{2}$$

Using $\{\hat{x}_{ij} \mid \hat{x}_{ij} = \mathbf{t}_i^{\mathrm{T}} \mathbf{v}_j\}$, matrix **W** and matrix **H** are obtained by repeating the following calculations until they converge.

$$t_{ik} \leftarrow t_{ik} \frac{\sum_j x_{ij} v_{kj}}{\sum_j \hat{x}_{ij} v_{kj}}, \quad v_{kj} \leftarrow v_{kj} \frac{\sum_i x_{ij} t_{ik}}{\sum_i \hat{x}_{ij} t_{ik}} \tag{3}$$

Please refer to the Appendix for the details on how Eq. 3 was derived.

Constructing trends of activities by place

By applying NMF, matrix **W** and matrix **H** are obtained. Matrix **H** has information about the time-series distributions of each activity type (the distributions correspond to the graphs in the right side of Fig. 3). The temporal resolution of the distributions is the given length of a time-slice. The time-series distribution of passengers for each station and each day is expressed as the linear sum of the time-series distributions of activity types. Matrix **W** has the information about the correlation to the basic distribution for each station and each day. Using the information, the trends of activity for each place are expressed. The temporal resolution of an activity trend is 1 day.

Let vector w_i denote the i-th column vector of **W**. The vector contains information on the volumes of visitors' arrivals for one travel activity across all places and during the entire duration of the study period. By taking elements of place s from vector w_i and sorting those elements in chronological order, series $x_{i,s} = \{a_j\}_{j=1}^{n} = \{a_1, .., a_j, .., a_n\}$ is obtained where n is the number of dates included in the data, and j denotes a day. This series represents the trend of the visitors' arrivals for trip activity i at place s.

Change point detection

In this section, we describe how EAT-CD finds the dates when the trend of the visitors' arrivals for each trip activity at each place has changed. Detecting a change point can present a challenge, as sudden changes may occur in a time-series distribution. Many efficient methods have been proposed (e.g. [34–36]). The objective of the studies is to find change points, and to develop an effective method to record the degree of change for each point. The degree of change can be derived by comparing partial distributions before and after the point.

The objective is to detect the change point from series $x_{i,s} = \{a_j\}_{j=1}^{n} = \{a_1, .., a_j, .., a_n\}$. Assuming that series $x_{i,s}$ draws a time-series distribution as shown in Fig. 6-(1), Day d_c is the date when a sudden change has occurred. The change point can be identified by investigating when the greatest degree of change is recorded for $j = c$.

The degree of change of point i is calculated by measuring the difference between series $\{a_j\}_{j=i-r}^{i-1} = \{a_{i-r}, .., a_{i-1}\}$ and series $\{a_j\}_{j=i}^{i-1+r} = \{a_i, .., a_{i-1+r}\}$ using Jensen-Shannon Divergence [37] (Fig. 6-(2)).

Jensen-Shannon divergence is a method used for measuring the similarity between two probability distributions. It is a modification of Kullback-Leibrer Divergence [38]. The Kullback-Leibler divergence from distribution Q to distribution P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{4}$$

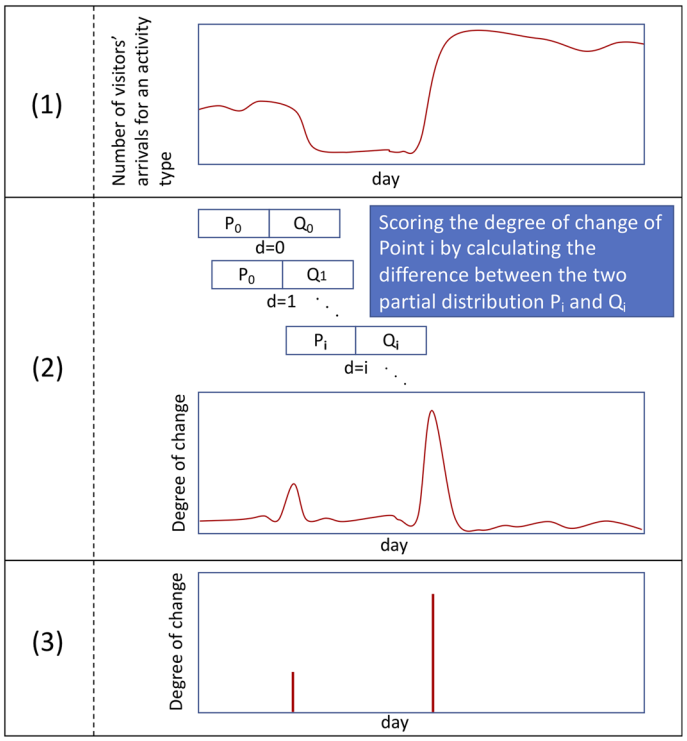


Fig. 6 Procedure to detect changes of the trend of activity in an area: (1) The proposed method detects changes of the inferred number of visitors' arrivals for each activity type. (2) The method scores the degree of change of point *i* by calculating the Jensen-Shannon divergence between the two partial distributions P and Q. (3) The method ignores the scores of points that are located near the maximum points

where p and q denote the probability density functions of P and Q, respectively. Assuming $P \sim N(\mu_1, \sigma_1^2)$ and $Q \sim N(\mu_2, \sigma_2^2)$,

$$D_{\mathrm{KL}}(P\|Q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{5}$$

Since this is asymmetric with respect to P and Q, Jensen-Shannon divergence is defined as follows to ensure symmetry.

$$D_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M) \tag{6}$$

where $M = \frac{P+Q}{2}$. Since M follows $M \sim N(\frac{(\mu_1+\mu_2)}{2}, \frac{\sigma_1^2+\sigma_2^2}{2})$,

$$D_{\mathrm{JS}}(P\|Q) = \frac{1}{2} \log \frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2} + \frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \tag{7}$$

Assuming that series $\{a_j\}_{j=i-r}^{i-1}$ and series $\{a_j\}_{j=i}^{i-1+r}$ follow the normal distribution, their variances and means are derived. The difference between series $\{a_j\}_{j=i-r}^{i-1}$ and series $\{a_j\}_{j=i}^{i-1+r}$ is calculated using Jensen-Shannon divergence. The score of change of point i is defined as follows:

$$S_i = D_{\mathrm{JS}}\left(\{a_j\}_{j=i-r}^{i-1} \middle\| \{a_j\}_{j=i}^{i-1+r}\right) \tag{8}$$

The score increases gradually before d_c and decreases gradually after d_c. Therefore, the exact change point can be identified by finding the peak value; i.e. if there is a point with a greater score than S_i in $[i-r, i+r]$, we ignore S_i (Fig.6-(3)). The score of change can be thus summarized as

$$S_i^* = \begin{cases} S_i & \left(\text{if } \; S_i = \max\limits_{i-r \le j \le i+r} S_j\right) \\ 0 & (\text{otherwise}) \end{cases} \tag{9}$$

Using this method, the change points for each place and each activity are extracted.

Assigning labels to activity patterns and reasons to detected changes

Activity patterns were not defined mechanistically in previous studies (Nishi et al. [23]; Fan et al. [26]). They identified each activity pattern by interpreting the spatio-temporal distributions of the activity pattern. Our study adopts a similar approach in that we did not design EAT-CD to extract and label each activity pattern. Rather, it is through the interpretation of the spatio-temporal patterns within the area that will inform us with the initial labelling of the activity types. In addition, while our method detects changes in the travel patterns, it is not intended as a means to explain reasons for such changes. In "Results and discussions" section, we will interpret the activity patterns extracted with EAT-CD and infer reasons of detected changes to better understand results.

Dataset

In principle, EAT-CD can be applied to any type of mobility data that have the hourly visitor count for each location/area, and it does not require detailed information on each trip.

To explore the validity of EAT-CD, this paper applies it to a set of smart card data from public transport in the Kansai Area, Japan. The Kansai Area is the second-most populated region in Japan after the Greater Tokyo Area. The data are collected at the auto fare collection barriers in each station. The original data have trip records with information on trip ID, passenger ID, boarding time, boarding station, alighting time, and alighting station. Data anonymisation is performed after extracting the data.

The time is discretised by dividing 1 day into n time-slices. In this experiment, we set the length of each time-slice to 1 h. From this data, we extract the hourly numbers of passengers' arrivals for each station, and we set the numbers to the values of $\mathbf{V}$. We use the information about alighting time and alighting station to extract hourly numbers of passengers' arrivals. We do not use information about trip ID, passenger ID, boarding time, or boarding station in this study. Since every train service ends before 2 a.m., we created daily temporal distribution of passengers' arrivals using data from 3 a.m. to 3 a.m. on the following morning. The original dataset was collected at 723 stations across 6 railway companies operating in the Kansai region, and the data were anonymised before they were shared with us. Figure 7 shows the locations of the stations. The data cover a 2-year period from April 2013 to March 2015. The average number of passengers is 1,087,351 per day, and the average number of trips is 2,477,966 per day.

Results and discussions

The daily temporal distributions of passengers' arrivals at all stations are decomposed, allowing us to detect changes in the patterns of activities at each station. This section shows the results of the application, namely the estimated break-down among different trip activities, followed by the activity trends for each station. The pattern of

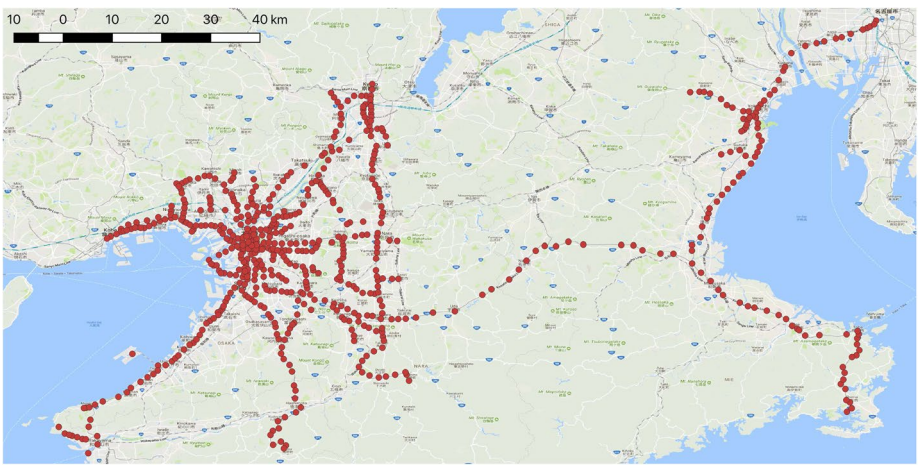


Fig. 7 The locations of the 723 railway stations used in this paper

each trip activity is labeled through the observation of the time-series distribution of each trip activity and the overall trends at the main terminal stations. For example, a sudden surge in the passenger volume around 8 a.m. can be classified as commute to work/school. Also, if there is a rise in the number of trips to the residential areas in the evening, they are regarded as the returning leg of commuters.

Basic distributions extracted by NMF

Using EAT-CD, the smart card data of public transport were classified into nine types of trip activities. The number of activity types may differ when this method is applied to mobility data in other areas or countries. Figure 8 shows the temporal distributions of these trip activities. EAT-CD enables us to extract basic distributions and construct activity trends of each station without relying the local land use data. However, identifying each activity type requires either a priori knowledge of people's life style in the region, or a heuristic process to interpret information about the region. In this study, we labeled each distribution as an individual activity by empirically interpreting the shape of the distribution. In this experiment, each distribution is labeled as one activity manually by interpreting the shape of the distribution.

In Fig. 8, trip Type 1 represents commute to work/school, showing a sudden increase in the volume during the rush hour at 8 am. Types 2 and 3 are labeled as returning home, which are dominant among stations in residential areas. Type 2 is particularly prevalent in those stations from Monday to Thursday, while Type 3 prevails on Friday. It is considered that workers usually go out for social activities on Friday because they do not need to wake up early the following morning. Types 4, 5, 6, and 7 are dominant among stations that serve as access nodes for concert halls and stadiums. They tend to exhibit a steep rise at a particular time of the day and are less periodic in their chronological patterns. These are labeled as events. The difference in the time of a steep rise reflects the beginning time of events. Types 8 and 9 are labeled as leisure activities. They show a milder change in their volume over the course of the day than Types 4, 5, 6, and 7 do and also tend to resonate with the change in the season. These include the viewing of cherry blossoms and autumn leaves, which are popular activities in Japan during April and September. For such events, visitors do not need to arrive at a specific time. Thus, forming a milder curve in their volume.

Trends of the activity types in the areas of stations

The trends of the trip activities near a station are extracted from $\mathbf{W}$. Each column vector of $\mathbf{W}$ contains information about the volume of the arriving passengers for each travel activity across all places for the entire duration of the study period. To extract the distribution of a single activity type for each station, one column vector of $\mathbf{W}$ is selected, then elements of the place are selected. We obtain the activity trend by sorting the selected elements in chronological order.

Figure 9 shows an illustrative example of the trends of the trip activities near a station that is located in a business area. In this area, Type 1 (commute to work/school) prevails as the dominant activity. The period between 11 August and 18 August falls on a Japanese Buddhists holiday week and shows a decline in the number of trip

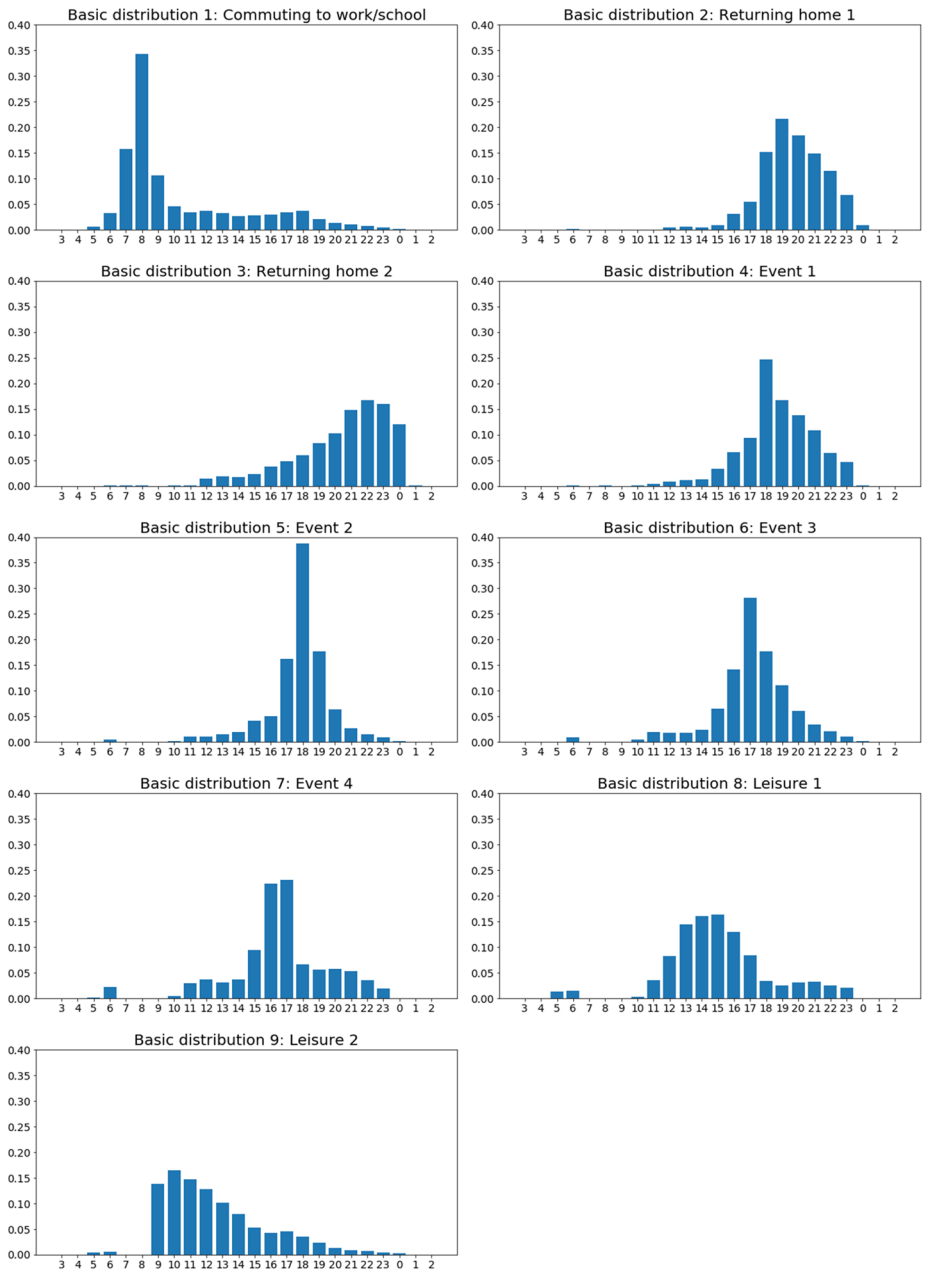


Fig. 8 The temporal distributions of the extracted activity types

activities. There is also a sudden increase for Type 5 trips (Event 2) on 25 July 2014. This is due to a famous annual firework festival held that day along the river near that station, attracting a large number of spectators.

Figure 10 shows another example of trip activities at a station in a residential area. In this area, Types 2 and 3 (returning home) are the dominant activities. On Fridays, the number of trips decreases for Type 2 and increases for Type 3. This is because people often go out for social activities and return home late on Fridays.

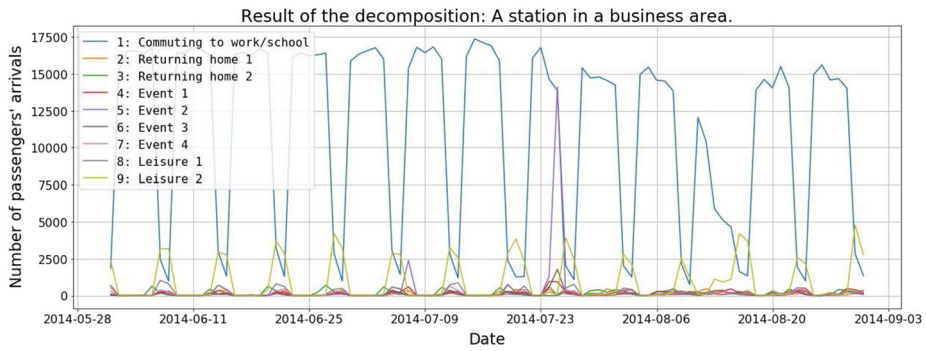


Fig. 9 The activity trend of a business area

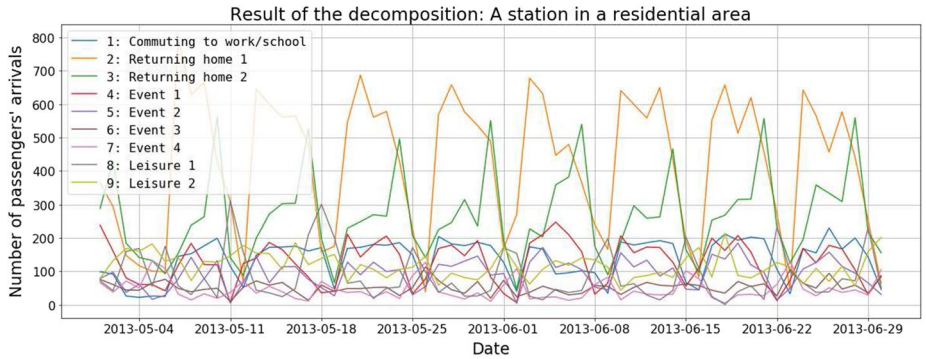


Fig. 10 The activity trend of a residential area

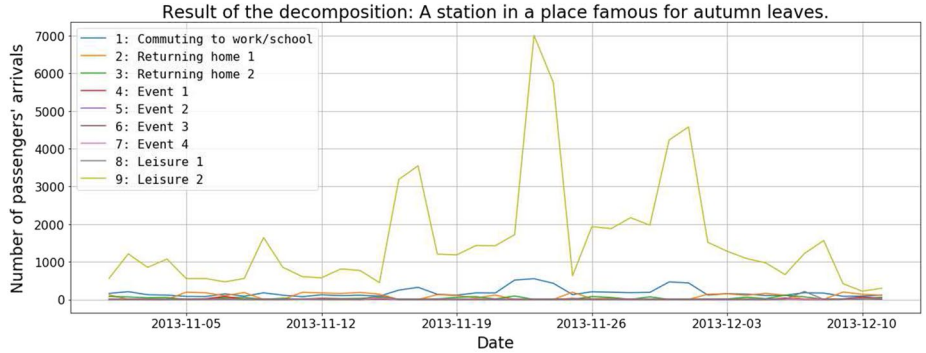


Fig. 11 The activity trend of an area famous for autumn leaves

Figure 11 shows the trends of trip activities at a station in an area known for the scenic beauty of autumn leaves. Viewing of autumn leaves is a popular activity in Japan. In this area, Type 9 trip (leisure 2) is the most dominant activity, especially on Saturday and Sunday when the number of Type 9 trips show a rapid increase.

Figure 12 shows the trends of trip activities at a station near a baseball stadium. In this area, Type 6 and 7 trips (events 3 and 4) are the most dominant activities. The dates when the trip activities suddenly increase coincide with the days of the baseball matches.

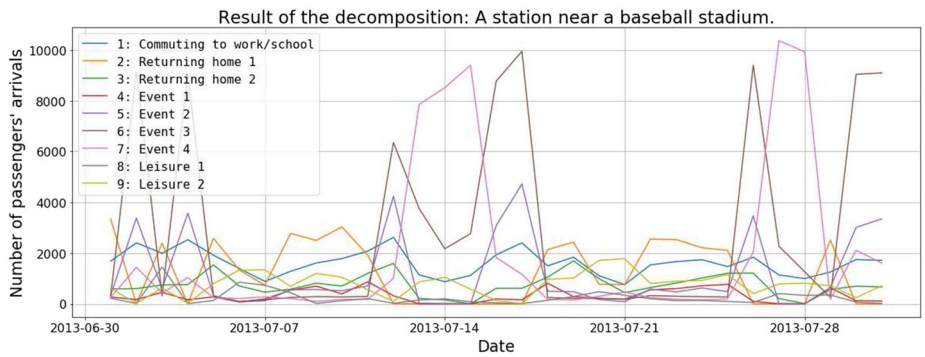


Fig. 12 The activity trend of an area with a baseball stadium

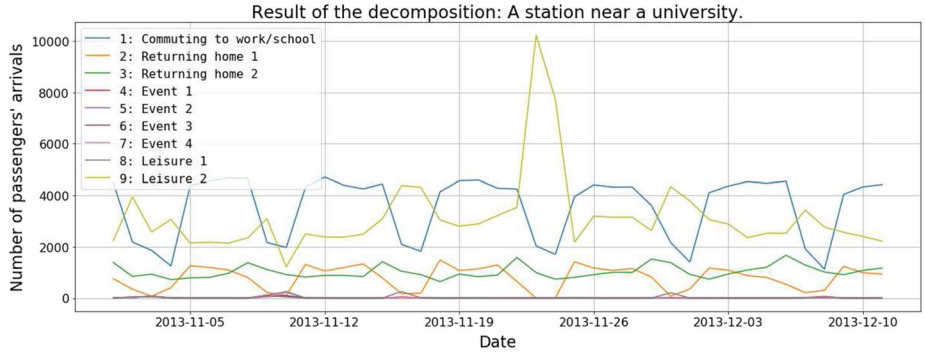


Fig. 13 The activity trend of an area where a university is located

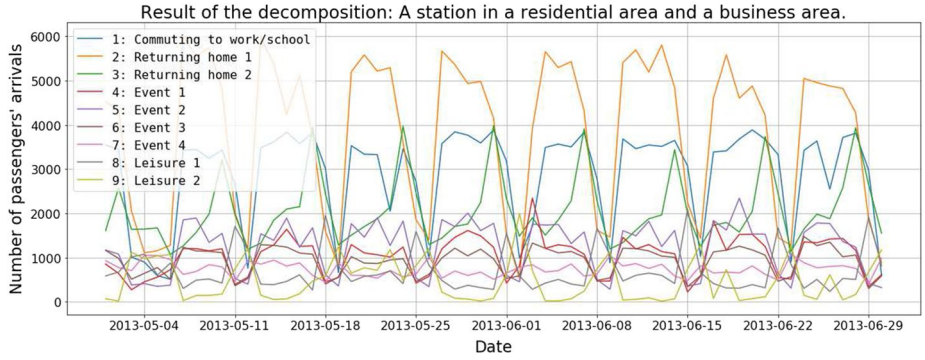


Fig. 14 The activity trend of an area of mixed land use for business and residence

Figure 13 shows the trends of trip activities at a station near a university. In this area, Type 1 trips (commute to work/school) mark the most dominant activity. Type 9 trips (Leisure 2) shows a one-off surge on 23 and 24 November when the university campus hosts a festival.

Figure 14 shows the trends of trip activities at a station in an area of mixed land use comprising business and residential functions. In this area, Types 1 (commute to work/

school), 2 (returning home 1), and 3 (returning home 2) are the most dominant activities, and the decomposition process is successful in capturing multiple activities in the area.

Evaluation of the results of the decomposition

Performance of EAT-CD is evaluated by comparing the results from the NMF with travel survey data collected by the Ministry of Land, Infrastructure, Transport and Tourism in Japan. The survey is carried out every 10 years. Respondents are given questionnaires, and they record travel histories on the questionnaires. Respondents record the destination, origin, departure time, arrival time, trip purpose, and the means of transportation of each trip. The detailed data is not open to the public, but aggregated data collected in the Kansai Area is available online [39]. We use the aggregated data that contains information on trip purpose, exit stations, classification of weekday, weekend and holiday, and the number of passengers. The items of trip purposes in the survey data are (1) returning home, (2) business, (3) private purposes, (4) commuting to work, and (5) commuting to school. We group information about trip purposes into three types: commuting to work/school, private purposes, and returning home. We aggregate travel survey data and the results of EAT-CD by redefining trip purposes listed in Table 2. The evaluation process was carried out separately for weekdays and weekends/holidays, and by different trip purposes. The numbers of arriving passengers at all stations recorded in the survey data are compared against the estimates obtained through our study. Pearson's correlation coefficient is calculated for measuring the accuracy of EAT-CD.

There is no travel survey data collected after 2010. Therefore, we compare the result of EAT-CD from smart card data collected in 2014 with travel survey data collected in 2010. Since those two types of data are from different periods, we consider that land use of some areas may have changed between the two periods. These differences reduce the correlation coefficients between the two types of data, and the correlation coefficients would increase if we could compare data collected in the same period. Therefore, we assume that it is sufficient if the correlation coefficient between the two types of data from different periods is high enough.

Table 3 shows the result of evaluation. As shown in this table, correlation coefficients are higher than 0.77, and p-values are lower than 6.54×10^{-96}. Therefore, we conclude

Table 2 Redefinition of trip purpose for evaluation

Redefined purposes	Travel survey data	EAT-CD
Commuting to work/school	(4) Commuting to work	Activity type 1 (Commuting to work/school)
	(5) Commuting to school	
Private purposes	(3) Private purposes	Activity type 4 (Event 1)
		Activity type 5 (Event 2)
		Activity type 6 (Event 3)
		Activity type 7 (Event 4)
		Activity type 8 (Leisure 1)
		Activity type 9 (Leisure 2)
Returning home	(1) Returning home	Activity type 2 (Returning home 1)
		Activity type 3 (Returning home 2)

Table 3 Comparison between the result of our method and the travel survey data

Purpose	Day of the week	Correlation coefficient	p-value
Commuting to work/school	Weekday	0.8775401	5.32×10^{-163}
Commuting to work/school	Holiday	0.7744606	6.54×10^{-96}
Private purposes	Weekday	0.8571198	6.93×10^{-148}
Private purposes	Holiday	0.9272756	1.78×10^{-218}
Returning home	Weekday	0.8124998	3.83×10^{-121}
Returning home	Holiday	0.8560062	5.86×10^{-148}

that the correlation coefficients are sufficiently high with low p-values to confirm that the outcome of EAT-CD sufficiently reflects the survey data.

Detecting changes over time

Table 4 shows the ranking of scores of change. Of the 15 highest scores, two of the nine activity types, namely Types 1 (commute to work/school) and 9 (leisure) are included. EAT-CD is designed for detecting changes in the activity trends, and it is not intended

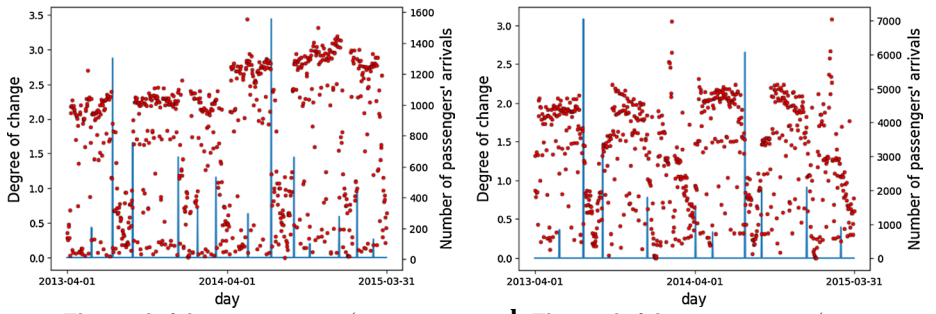
Table 4 Ranking of scores of change. The top fifteen highest scores are dominated by Types 1 (commute to work/school) and 9 (leisure) activities only, which confirms the abrupt nature of the changes in the two types of trip activities

Rank	Score of change	Activity type	Station ID	Characteristic of the area	Date	Reason for the change
1	3.439454972	1	Station A	Universities and high schools	2014-07-13	Ending of semester
2	3.081841026	1	Station B	Universities and high schools	2013-07-23	Ending of semester
3	3.041180957	1	Station C	Universities and high schools	2014-07-18	Ending of semester
4	2.874711479	1	Station A	Universities and high schools	2013-07-16	Ending of semester
5	2.829092656	1	Station D	Universities and high schools	2014-07-22	Ending of semester
6	2.788287796	1	Station E	Universities and high schools	2013-07-14	Ending of semester
7	2.756890751	9	Station F	Shopping center	2014-03-11	Extension of a shopping center
8	2.737122907	1	Station E	Universities and high schools	2014-09-03	Beginning of semester
9	2.724598206	1	Station G	Universities and high schools	2014-09-03	Beginning of semester
10	2.717382240	1	Station H	Universities and high schools	2014-08-28	Beginning of semester
11	2.653081428	1	Station B	Universities and high schools	2014-07-27	Ending of semester
12	2.503246179	9	Station I	Shopping center	2014-03-15	Opening of a shopping center
13	2.366644053	1	Station C	Universities and high schools	2013-07-16	Ending of semester
14	2.341568280	1	Station D	Universities and high schools	2013-07-20	Ending of semester
15	2.196827999	1	Station E	Universities and high schools	2014-07-12	Ending of semester

as a means to infer the reasons for such changes. We infer the reasons for the changes by the neighborhood characteristics for better understanding of the results. Thirteen changes in Activity Type 1 are caused by the beginnings or endings of academic semesters. Two changes in Activity Type 9 are caused by the development of a new shopping center and extension of an existing centre, respectively.

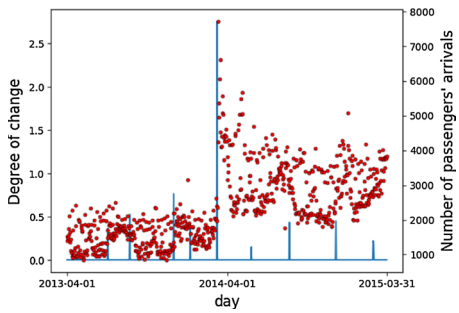
Figure 15a–d shows the trends reflecting the 1st, 2nd, 7th and 12th changes respectively from Table 4 and their degrees of change. The red points illustrate the inferred numbers of passengers' arrivals for the Type 1 (Fig. 15a, b) and Type 9 (Fig. 15c, d). As shown in Fig. 15a, b, seasonal changes linked to academic semesters are detected, while non-seasonal changes are also detected (Fig. 15c, d). Figure 16 shows the graph of the trend of activity Type 9 (leisure) of Station F for a short period (1 month) near the detected change. The shopping center reopened on March 12, 2014. However, some shops started one day before the official reopening of the shopping center. EAT-CD captures the change on March 11, 2014.

Figure 17 shows the trend of all trip activities in the area whose change is ranked 12th in Table 4. In this area, a large shopping center opened. The dashed red line indicates the opening date of the shopping center (March 16, 2014). The number of passengers commuting to work/school rapidly increases 2 days before the opening dates but sees a sharp

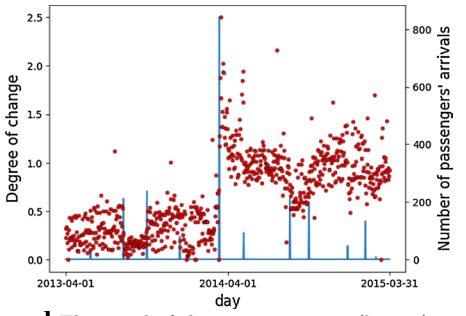


a The trend of the activity type 1 (commuting to work/school) of Station A: An area of universities and high schools.

b The trend of the activity type 1 (commuting to work/school) of Station B: An area of universities and high schools.



c The trend of the activity type 9 (leisure) of Station F: An area of a newly extended shopping center.



d The trend of the activity type 9 (leisure) of Station I: An area of a newly opened shopping center.

Fig. 15 Detected changes in activity trends: The blue-colored graphs are the degrees of change. The red points are the inferred number of visitors' arrivals for commuting to work/school (**a**, **b**) and leisure (**c**, **d**). In areas of universities and high schools, changes are detected from the trends of the activity type 1 (commuting to work/school) according to the beginnings and endings of semesters. In areas of shopping centers, changes are detected from the trends of the activity type 9 (leisure) according to the opening and extension of the shopping centers

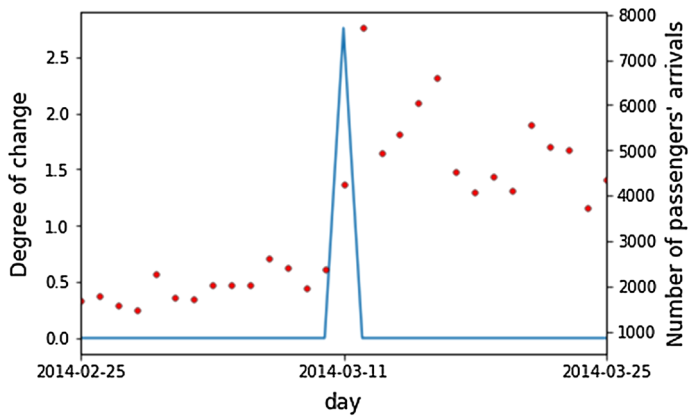


Fig. 16 The graph of the trend of the activity type 9 (leisure) of Station F in a short period (1 month) near the detected change: an area of a newly extended shopping center

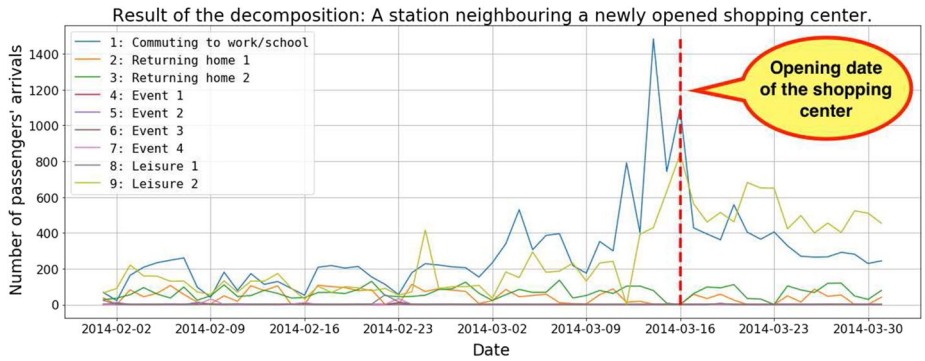


Fig. 17 The activity trends of the area of a newly opened shopping center

decline after the opening dates. On the other hand, the number of passengers for leisure (Type 9) increases on the opening date and remains higher than it was before the opening date. It is considered that the increase of commuters is caused by the opening staff and temporary helpers for the launch of the shopping center, whereas the consistently high number of passengers for leisure purpose confirms that the shopping center continues to attract people after the opening date. However, without showing the break-down by trip purposes, the change in the total number of arriving passengers could mislead us to think that the shopping center attracted people on the opening date only.

In summary, the results confirmed that EAT-CD detects urban changes by constructing activity trends and scoring the degrees of change. Chronological change in each activity trend reflects such urban changes (Fig. 15). Scoring the degree of changes succeeds in capturing the exact day when urban change has occurred (Fig. 16). In addition, it is possible to see the impact of an urban change on each activity type from the activity trends of period near detected changes (Fig. 17).

We note that the change detection would be more efficient if similar activities (e.g. returning home 1 and returning home 2) were combined before measuring the extent of changes, as we have combined them for comparing the results of EAT-CD with travel survey data (Table 2). However, introduction of descriptive labels requires the

interpretation of extracted activity types, and they may vary for mobility data collected in other areas or countries.

The significance of the proposed method in the context of studies on inference of land use/activity

The proposed method (EAT-CD) succeeds in addressing the following points that the previous studies have not:

- The method extracts the temporal distributions of activity types.
- The method uses only hourly numbers of visitors' arrivals in each area.
- The method decomposes the number of visitors' arrivals in each area and each day into the activity types, and it constructs the trends of the activity types for each area.
- The method scores the degree of changes for each activity and each area, and it detects significant changes.
- It is possible to analyze the causes of changes by looking at the trends of the activity types.

Compared to other existing methods such as those proposed by Nishi et al. [23] and Fan et al. [26], EAT-CD has the following advantages: First, EAT-CD captures mixed land use as shown in Fig. 14. Second, it captures area-specific changes as shown in Fig. 15.

With some adjustment, EAT-CD can be also applied to other types of locational data such as GPS data collected from mobile phones. Nishi et al. [23] and Fan et al. [26] analyse time-series distributions of people's visits using GPS data collected from mobile phones. To make time-series distributions of people's visits, they discretise the time and coordinates. They divide the study area of a city into grid cells and the duration of 1 day into time-slices. The number of human inflows to each tile based on human coordinates is used to create time-series distributions of visitors. By using the numbers, it is possible to perform EAT-CD for GPS data collected from mobile phones. Since EAT-CD can automatically detect urban changes by using mobility data such as smart card data of public transportation and GPS data collected from mobile phones, it will be possible to monitor urban changes globally. This is significantly advantageous for urban planners.

Conclusion

In this paper, we proposed a method called EAT-CD to decompose the number of passengers arriving at each station into activity types, to construct activity trends for each area, and to detect changes in the land use patterns of urban areas. On **RQ1**, we confirmed that the numbers of visitors to each place can be estimated for different activity types by applying NMF to the hourly numbers of visitors recorded in the mobility data. On **RQ2**, we confirmed that a method can be developed for detecting changes in the activity trends using the estimated numbers of arriving passengers for each trip activity by using Jensen-Shannon divergence. EAT-CD only requires the total number of passengers by the hour. Validity of EAT-CD was examined by applying it to a set of smart card data from public transport in the Kansai Region, Japan. The results showed that the activity trends were successfully derived and significant changes in the patterns of travel or land use were well detected.

We note limitations of our work. The number of activity types is determined manually in this study, and labeling of activity types is also conducted manually. Decomposing the number of visitors and detecting changes requires no prior knowledge of the region or the subjective setting of the trip types. However, labeling each activity type and finding the reasons for changes requires prior knowledge. It is necessary to modify the method to automatically execute the procedure.

Another limitation of our study is the difficulty in evaluating the outcomes. While the breakdown of visitor numbers by the respective activity type has been compared with travel survey data, the results from the change detection was never assessed. This is because the travel survey data contains no chronological information on passengers' movements. Identifying suitable dataset with multiple timepoints would help examine the sensitivity of change detection. Another item for our research agenda would be a systematic comparison of the performance of EAT-CD with that of other existing methods. EAT-CD offers two different types of outcomes, namely the decomposition of population into activity types and the detection of changes in the volume of each activity type. There are several methods that allow us to decompose population into different activity types and also detect changes in their volume. Their performance can be compared by either (1) using empirical data that have information about the number of visitors from multiple time points; or (2) conducting simulation-based analysis of human movements. For example, Fan et al. [26] simulate human movements with respect to the spatial distribution of the Points of Interest (POIs).

EAT-CD can be applied to any kinds of data that have information about time-series distribution of visitors and would prove beneficial to planners who work on developing cities.

The ultimate goal of our research is to develop a method to capture urban changes that cannot be detected by other means. EAT-CD can detect changes that are known and established (e.g. the start and end of a semester, and opening of a shopping center), as well as those arising from previously unknown factors. On the other hand, EAT-CD has succeeded in capturing the effect caused by such big changes. We have shown that EAT-CD is capable of capturing changes in the number of visitors for leisure purposes to the station close to a newly opened shopping center and that this increase stays on where the number of passengers remain higher than it was before the opening (Fig. 17). The main advantage of EAT-CD is that it can capture changes that cannot be captured without decomposing the number of visitors into activity types. We expect that it is possible to develop a method to detect and understand the gradual changes of urban characteristics using EAT-CD. Our future work will be dedicated to realising such development of EAT-CD.

Our future work will include the implementation of systems to apply EAT-CD to various kinds of mobility data to monitor urban changes in various areas. Processing a large amount of mobility data and automatic labeling of activity types and detected changes are main challenges. To address the challenges, further studies are needed employing ever developing methodologies of machine learning and urban studies.

Abbreviations

API: application programming interface; CDR: call detail records; CHMM: continuous hidden Markov model; DBSCAN: density-based spatial clustering of applications with noise; DP: Dirichlet process; EAT-CD: extraction of activity types and

change detection; EM: expectation-maximization; GMM: Gaussian mixture model; GPS: global positioning system; NMF: non-negative matrix factorization; NTF: non-negative tensor factorization; OD: origin-destination; POI: point of interest; SVM: support vector machine; TfL: Transport for London.

Authors' contributions
All the authors discussed and designed the experiments as well as contributing to the writing of the paper. TN.M defined the research agenda, implemented the experiments and wrote the manuscript. All the authors read and approved the final manuscript.

Author details
[1] The University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo 113-8656, Japan. [2] King's College London, Strand, London WC2R 2LS, UK.

Acknowledgements
This study was mainly carried out while T. N. Maeda was studying as a visiting research student under the supervision of N. Shiode and C. Zhong at King's College London. The visit was supported by the Leading Graduates Schools Program "Global Leader Program for Social Design and Management (GSDM)" run by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Funding
Not applicable.

Appendix: Detailed explanation about deriving the algorithm of NMF

The procedure to introduce Eq. 3 is as follows:

Equation 10 is obtained by omitting constants from Eq. 1.

$$F_{\mathrm{Eu}}(\mathbf{T},\mathbf{V}) = \sum_{i,j}\left[(t_{ik}v_{kj})^2 - 2x_{ij}\mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j\right] \tag{10}$$

An auxiliary function, $F^{+}_{\mathrm{Eu}}(T,V,R)$ is defined by adding auxiliary variables r_{ijk} that satisfy Eq. 12.

$$F^{+}_{\mathrm{Eu}}(\mathbf{T},\mathbf{V},\mathbf{R}) = \sum_{i,j}\left[\sum_{k}\frac{(t_{ik}v_{kj})^2}{r_{ijk}} - 2x_{ij}\mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j\right] \tag{11}$$

$$r_{ijk} > 0,\ \sum_{k=1}^{K} r_{ijk} = 1 \tag{12}$$

Function F^{+}_{Eu} satisfies Eqs. 13 and 14.

$$F_{\mathrm{Eu}}(\mathbf{T},\mathbf{V}) \le F^{+}_{\mathrm{Eu}}(\mathbf{T},\mathbf{V},\mathbf{R}) \tag{13}$$

$$F_{\mathrm{Eu}}(\mathbf{T},\mathbf{V}) = \min_{\mathbf{R}} F^{+}_{\mathrm{Eu}}(\mathbf{T},\mathbf{V},\mathbf{R}) \tag{14}$$

Eqs. 15, 16, 17, and 18 are derived by using Lagrange multipliers method.

$$L(\mathbf{T},\mathbf{V},\mathbf{R},\mathbf{\Lambda}) = F^{+}_{\mathrm{Eu}} + \sum_{i,j}\lambda_{ij}\left(\sum_{k} r_{ijk} - 1\right) \tag{15}$$

$$\frac{\partial L}{\partial r_{ijk}} = -\frac{(t_{ik}v_{kj})^2}{r_{ijk}^2} + \lambda_{ij} = 0 \tag{16}$$

$$\frac{\partial F_{\mathrm{Eu}}^{+}}{\partial t_{ik}} = 2t_{ik}\sum_j \frac{v_{kj}^2}{r_{ijk}} + 2\sum_j x_{ij}v_{kj} = 0 \tag{17}$$

$$\frac{\partial F_{\mathrm{Eu}}^{+}}{\partial v_{kj}} = 2v_{kj}\sum_i \frac{t_{ik}^2}{r_{ijk}} + 2\sum_i x_{ij}t_{ik} = 0 \tag{18}$$

Eq. 19 is derived from Eqs. 12 and 16.

$$r_{ijk} = \frac{t_{ik}v_{kj}}{\mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j} = \frac{t_{ik}v_{kj}}{\hat{x}_{ij}} \tag{19}$$

Eq. 20 is derived from Eqs. 17 and 18.

$$t_{ik} = \frac{\sum_j x_{ij}v_{kj}}{\sum_j \frac{v_{kj}^2}{r_{ijk}}},\quad v_{kj} = \frac{\sum_i x_{ij}t_{ik}}{\sum_i \frac{t_{ik}^2}{r_{ijk}}} \tag{20}$$

Finally, Eq. 3 is derived from Eqs. 19 and 20.

Publisher's Note

Received: 7 November 2018 Accepted: 3 January 2019
Published online: 14 January 2019

References

1. Jones PM. New approaches to understanding travel behaviour: the human activity approach. Oxford: Oxford University; 1977.
2. Axhausen KW, Zimmermann A, Schönfelder S, Rindsfüser G, Haupt T. Observing the rhythms of daily life: a six-week travel diary. Transportation. 2002;29(2):95–124.
3. Zheng Y, Capra L, Wolfson O, Yang H. Urban computing: concepts, methodologies, and applications. ACM Trans Intell Syst Technol. 2014;5(3):38–13855.
4. World Bank. Big data in action for development. Washington, DC: World Bank; 2014.
5. United Nations Global Pulse. Big data for development: challenges and opportunities. New York: UN Global Pulse; 2012.
6. United Nations Global Pulse. Mobile phone network data for development. New York: UN Global Pulse; 2013.
7. Smith C, Quercia D, Capra L. Finger on the pulse: identifying deprivation using transit flow analysis. In: Proceedings of the 2013 conference on computer supported cooperative work. CSCW '13. New York: ACM. 2013, p. 683–692.
8. Blumenstock JE. Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. Inf Technol Dev. 2012;18(2):107–25.
9. Alexander L, Jiang S, Murga M, González MC. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp Res Part C Emerg Technol. 2015;58:240–50.
10. Han G, Sohn K. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden markov model. Transp Res Part B Methodol. 2016;83:121–35.
11. Lee SG, Hickman M. Trip purpose inference using automated fare collection data. Public Transp. 2014;6(1):1–20.
12. Alsger A, Tavassoli A, Mesbah M, Ferreira L, Hickman M. Public transport trip purpose inference using smart card fare data. Transp Res Part C Emerg Technol. 2018;87:123–37.
13. Hu N, Legara EF, Lee KK, Hung GG, Monterola C. Impacts of land use and amenities on public transport use, urban planning and design. Land Use Policy. 2016;57:356–67.

14. Wang P, Fu Y, Liu G, Hu W, Aggarwal C. Human mobility synchronization and trip purpose detection with mixture of hawkes processes. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '17. New York: ACM .2017 , p. 495–503.
15. Yao D, Yu C, Jin H, Ding Q. Human mobility synthesis using matrix and tensor factorizations. Inf Fusion. 2015;23:25–32.
16. Zhou Y, Fang Z, Zhan Q, Huang Y, Fu X. Inferring social functions available in the metro station area from passengers' staying activities in smart card data. ISPRS Int J Geo Inf. 2017;6(12):394.
17. Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: a large-scale application in the netherlands. Transp Res Part C Emerg Technol. 2009;17(3):285–97.
18. Wang J, Kong X, Rahim A, Xia F, Tolba A, Al-Makhadmeh Z. Is2fun: identification of subway station functions using massive urban data. IEEE Access. 2017;5:27103–13.
19. Xiao G, Juan Z, Zhang C. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transp Res Part C Emerg Technol. 2016;71:447–63.
20. Zhong C, Huang X, Arisona SM, Schmitt G, Batty M. Inferring building functions from a probabilistic model using public transportation data. Comput Environ Urban Syst. 2014;48:124–37.
21. Le Q, Mikolov T. Distributed representations of sentences and documents. In: International conference on machine learning. 2014, p. 1188–1196.
22. Long Y, Thill J-C. Combining smart card data and household travel survey to analyze jobs–housing relationships in beijing. Comput Environ Urban Syst. 2015;53:19–35.
23. Nishi K, Tsubouchi K, Shimosaka M. Extracting land-use patterns using location data from smartphones. In: Proceedings of the first international conference on IoT in urban space. URB-IOT '14. 2014, p. 38–43.
24. Frias-Martinez V, Frias-Martinez E. Spectral clustering for sensing urban land use using twitter activity. Eng Appl Artif Intell. 2014;35:237–45.
25. Chen Y, Liu X, Li X, Liu X, Yao Y, Hu G, Xu X, Pei F. Delineating urban functional areas with building-level social media data: a dynamic time warping (dtw) distance based k-medoids method. Landsc Urban Plan. 2017;160:48–60.
26. Fan Z, Song X, Shibasaki R. Cityspectrum: a non-negative tensor factorization approach. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing. UbiComp '14. New York: ACM. 2014, p. 213–223.
27. Cichocki A, Zdunek R, Phan AH, Amari S-i. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. New Jersey: John Wiley & Sons; 2009.
28. Wang J, Gao F, Cui P, Li C, Xiong Z. Discovering urban spatio-temporal structure from time-evolving traffic networks. In: Chen L, Jia Y, Sellis T, Liu G, editors. Web technologies and applications. Cham: Springer; 2014. p. 93–104.
29. Zhong C, Arisona SM, Huang X, Batty M, Schmitt G. Detecting the dynamics of urban structure through spatial network analysis. Int J Geogr Inf Sci. 2014;28(11):2178–99.
30. Zhong C, Schläpfer M, Arisona SM, Batty M, Ratti C, Schmitt G. Revealing centrality in the spatial structure of cities from human activity patterns. Urban Stud. 2017;54(2):437–55.
31. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Leen TK, Dietterich TG, Tresp V, editors. Advances in neural information processing systems 13. Massachusetts: MIT Press; 2001. p. 556–62.
32. Sawada H, Kameoka H, Araki S, Ueda N. Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans Audio Speech Lang Process. 2013;21(5):971–82.
33. Févotte C, Bertin N, Durrieu J-L. Nonnegative matrix factorization with the itakura-saito divergence: with application to music analysis. Neural Comput. 2009;21(3):793–830.
34. Takeuchi J, Yamanishi K. A unifying framework for detecting outliers and change points from time series. IEEE Trans Knowl Data Eng. 2006;18(4):482–92.
35. Kawahara Y, Sugiyama M. Sequential change-point detection based on direct density-ratio estimation. Stat Anal Data Min. 2012;5(2):114–27.
36. Matteson DS, James NA. A nonparametric approach for multiple change point analysis of multivariate data. J Am Stat Assoc. 2014;109(505):334–45.
37. Lin J. Divergence measures based on the shannon entropy. IEEE Trans Inf Theory. 1991;37(1):145–51.
38. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22(1):79–86.
39. The Ministry of Land, Infrastructure, Transport and Tourism in Japan. Person trip data in the western area in Japan (Written in Japanese), https://www.kkr.mlit.go.jp/plan/pt/index.html. Accessed 12 Sept 2018.