Information Sciences 607 (2022) 1311–1327

Contents lists available at ScienceDirect

Information Sciences

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

An unsupervised approach for semantic place annotation of trajectories based on the prior probability

Junyi Cheng [a], Xianfeng Zhang [a,*], Peng Luo [b], Jie Huang [a], Jianfeng Huang [a]

[a] Institute of Remote Sensing and Geographic Information Systems, Peking University, 5 Summer Palace Road, Beijing 100871, China
[b] Chair of Cartography, Technical University of Munich, 80333 Munich, Germany

ARTICLE INFO

Article history:
Received 27 September 2021
Received in revised form 5 June 2022
Accepted 10 June 2022
Available online 15 June 2022

Keywords:
Trajectory data mining
Semantic annotation
Place matching
Location-aware computing
Spatiotemporal probability

ABSTRACT

Semantic place annotation can provide individual semantics, greatly helping the field of trajectory data mining. Most existing methods rely on annotated or external data and require retraining models following a region change, thus preventing their large-scale applications. Herein, we propose an unsupervised method denoted as UPAPP for the semantic place annotation of individual trajectories using spatiotemporal information. The Bayesian Criterion is specifically employed to decompose the spatiotemporal probability of visiting the candidate place into spatial probability, duration probability, and visiting time probability. Spatial information in two geospatial data sources is comprehensively integrated to calculate the spatial probability. In terms of the temporal probabilities, the Term Frequency–Inverse Document Frequency weighting algorithm is used to count the potential visits to different place types in the trajectories and to generate the prior probabilities of the visiting time and duration. Finally, the spatiotemporal probability of the candidate place is then combined with the importance of the place category to annotate the visited places. Experimental results in a trajectory dataset collected by 709 volunteers in Beijing showed that our method achieved an overall and average accuracy of 0.712 and 0.720, respectively, indicating that the visited places can be annotated accurately without any annotated data.

1. Introduction

Semantic place annotation refers to determining the most likely visited places hidden in the raw trajectory data by contextual information (e.g., spatial and temporal information). With the increasing popularity of mobile terminals and the development of positioning technology, massive amounts of human movement trajectory data have been collected and recorded. Many studies have investigated various spatiotemporal patterns in the trajectory data [1,2], whereby a spatiotemporal trajectory is typically expressed as a series of coordinates, clusters or geographic regions. Such a large number of spatiotemporal trajectories provides us with new opportunities to discover the purposes and habits of individuals' travel patterns. However, the absence of semantic information related to the visited places prevents a comprehensive understanding of the daily behavior and complicates its subsequent usage. Understanding the semantics contained in the massive spatiotemporal trajectory data is of great significance for personalized service recommendations and predictions, the advance prediction and prevention of emergencies, and personal health monitoring [3,4]. Therefore, there is a growing demand for

* Corresponding author.
E-mail address: xfzhang@pku.edu.cn (X. Zhang).

https://doi.org/10.1016/j.ins.2022.06.034
0020-0255/

the automated semantic annotation of the trajectories before mining valuable travel patterns hidden in the huge spatiotemporal trajectory data [5–9].

Existing methods for semantic place annotation typically include two steps, namely, trajectory segmentation and semantic annotation. The process of subdividing the original trajectory into sequences in which "moving" and "stop" episodes alternately appear is called trajectory segmentation [10]. In particular, a "stop" is not entirely stationary, rather it denotes staying in a relatively small neighborhood for an extended period. It is assumed that if individuals have stopped, they must be doing some meaningful activities at specific places, and therefore the stops are noteworthy [11]. Semantic annotation is also a semantic enrichment process applied by integrating different types of data to obtain beneficial information from the analysis of a raw trajectory. The annotation of the moving episode typically includes inferring the travel mode, while the annotation of the stop episode determines the most probable places visited by the individual in their stops. Compared with moving episodes, stops exhibit richer semantic information and are considered more important in trajectory data mining. However, the accuracy of place annotation is limited by the following factors: 1) the positioning terminal often has low accuracy and cannot effectively match the actual location; 2) place ambiguity is a common occurrence, particularly for multiple place types in the same area (e.g., multiple shops or restaurants in a shopping mall); and 3) temporal ambiguity is also observed, for example, visiting the same place for different purposes results in complicated temporal patterns.

Many studies have recently been conducted on the semantic place annotation of trajectories. These approaches typically aim to determine the most likely visited places hidden in the trajectory data by multi-source information. However, they either ignore the temporal information of the stop or rely on annotated samples or other external data [6,8]. Moreover, obtaining large-scale annotated data usually proves to be difficult. Therefore, a semantic annotation model that can effectively combine temporal and spatial information without supplementary data is urgently required.

In this study, we develop a prior probability-based method for unsupervised semantic annotation of trajectory stop places. The contribution of our work is as follows:

1) A prior probability statistical method for semantic trajectory annotation that does not require annotated data or other external data is proposed. With only the trajectory data, our method can extract the visiting time and duration patterns for the different types of visited places. It begins with the trajectory dataset itself and is not restricted by changing regions and people. Thus, our method solves the dependence on annotated data in the existing methods and can greatly improve the availability of trajectory data.
2) A probability model that fully considers the spatial characteristics, visiting time and duration of the stop episode is created to address the issue of place ambiguity. Our model is able to comprehensively employ temporal characteristics and spatial locations to calculate the spatiotemporal probability of visiting different places, combining the spatiotemporal probability of the place and the importance of the place category to annotate the visited place.
3) A method that determines the spatial probability by combining the ROI (region of interest) and POI (point of interest) data is proposed to increase the accuracy of the spatial probability calculations, fully employing the topological characteristics provided by the ROI data and the spatial distribution characteristics of the POI.

2. Related work

2.1. Trajectory segmentation

Current trajectory segmentation methods can be classified into supervised, unsupervised, and semi-supervised approaches [12,13]. The supervised method uses labeled samples for training and learns the knowledge to generate sub-trajectories (e.g., the recently proposed WS-II) [14]. The semi-supervised approach includes RGRASP-SemTS [15] and CoEx-DBSCAN [16] and uses a combination of user-labeled meaningful segments and unlabeled data to segment the trajectory. Unfortunately, most trajectory datasets do not contain semantic labels, and it is challenging to label big trajectory datasets manually. Consequently, the unsupervised method is an essential topic in trajectory segmentation.

Unsupervised segmentation algorithms can be divided into four categories, namely, rule-based, clustering-based, sliding-window-based, and cost-function-based approaches. The rule-based algorithm detects the trajectory sequence that meets the required spatiotemporal conditions via a threshold. Commonly used thresholds include spatial distance, time length [17], speed [18], and angle change thresholds [19]. Although a rule-based method is easy to implement and relatively intuitive, it is not robust to noise. The cost-function-based approach partitions a trajectory by minimizing a specific cost function to build the most homogeneous segments, such as GRASP-UTS and TS-MF based on the minimum description length principle [12,20]. The sliding window-based approach determines where the moving object changed its behavior by deriving the local features within a fixed-size sliding window (e.g., OWS [21] and SWS [13]). While these two methods can effectively detect changes in the motion state, they are not suitable for the semantic mining of visited places. This is because an individual is not necessarily completely static when visiting a place such as walking in a park.

A major application of clustering-based segmentation is the detection of stop-and-move patterns [22], which describe the behavior of an individual that stays within a specific region for more than a period of time. Clustering-based methods detect stops by identifying continuous sequences adjacent in space and time. Moreover, these methods redefine the concepts of neighborhood and density to find points that are nearby in both space and time. CB-SMOT is a key clustering-based algorithm that redefines the neighborhood concept in DBSCAN, taking the longest continuous sequence with an average speed

less than the speed threshold as the neighborhood [23]. If the continuous sequence time exceeds the time threshold, it is considered a stop. Otherwise, it is considered to be a moving episode. However, the methods may introduce extra moving points at the beginning and end of the extracted point sequence. To overcome this weakness, TrajDBSCAN is proposed and assumes that the distances to the core point from all points in the neighborhood should be less than the distance threshold [24]. In addition, the time length of the sequence is used to replace the neighborhood density in DBSCAN. This allows for the application of TrajDBSCAN to trajectories of different sampling intervals. Subsequent work has been developed using this clustering-based concept [25–27]. Such methods are data-driven and relatively robust against noise. Therefore, in the current paper, a clustering-based method is adopted for the extraction of the stop episodes and to segment the trajectory data.

2.2. Semantic annotation

The semantic enrichment process annotates semantic information in trajectories fused with different sources of information. Hence, different semantic trajectory models with semantic web standards have been proposed in order to represent trajectory episodes and contextual information, such as STEP [28], FrameSTEP [29], Master [30], SEMANTIC-SEG [8] and SEPSIM [31]. This method can apply reliable external resources (e.g., a geographic knowledge base) and movement features to label the trajectories in question. However, such studies are limited to a basic reasoning mechanism using spatiotemporal concepts, properties, or relationships [31]. Due to temporal and spatial ambiguity, annotating visited places hidden in the trajectory is a complicated task. In the context of trajectory data mining, this work positions itself in the reasoning stage.

In order to solve place ambiguity, the features extracted from multi-source data can be comprehensively utilized to infer the places most likely to be visited. Commonly used features can be divided into geographic-, temporal-, and individual-related features. The features related to geographic data include spatial topological relationships [11,32], distance [33], and the distribution of place types [5,7]. Time-related features typically include the start time and duration of stop episodes [8,34] and activity history-related features, such as accumulated duration and frequency. Common individual-related features compromise the individual's occupation, age and home address [35]. Additional information is also selected, such as place popularity and moving patterns typically obtained from social media and review sites [8,36,37]. Individual-related features are difficult to obtain and are guarded by privacy restrictions. Thus, they are rarely employed in large-scale applications. Furthermore, the distribution of samples obtained from social media is often biased, and retraining is required following a region change. Such information is also relatively difficult to acquire. Geographic and temporal information, which can be obtained easily and is directly related to the travel patterns of individuals, is widely used in previous studies.

On the other hand, current methods for the reasoning of visited places can be grouped into rule-based, machine learning-based, and probability-based approaches. The rule-based method annotates visited places by formulating temporal and spatial rules [38]. For example, the most commonly used spatial rule takes the nearest POI as the visited place [39]. However, the rules are relatively arbitrary, preventing the place and temporal ambiguity from being solved effectively. The machine learning-based method uses labeled data for supervised learning and employs the trained model to annotate the visited places. Commonly used machine learning models in semantic annotation include decision trees, random forests [33,34], and neural networks [40]. A recent study established a random forest model based on temporal, spatial and sequential features for automatically estimating the semantic meanings of personal locations [7]. This method can employ a variety of information and is suitable for large-scale datasets, yet it relies on labeled data for training [35,40–42], and retraining is required after the region is changed. The probability-based method calculates the probability of visiting each place under the corresponding spatial location and time conditions. The majority of studies using the probability-based method employ spatial probability. For example, Yan established a spatial probability calculation model based on a two-dimensional Gaussian distribution function [5]. Many researchers subsequently followed a similar approach to this spatial probability model [6,8]. Unlike the aforementioned methods, some recent studies have adopted ROI data to establish a spatial probability model based on spatial topological relationships [11]. Current studies use either the ROI or POI when establishing the spatial probability model, while research that comprehensively integrates information from different types of geographic data is lacking. For the formulation of the temporal probability, previous studies generally directly calculate the probability from the activity log or employ social media to mine the required temporal patterns. For example, Gong calculated the visiting

Table 1
A summary of the differences between our method and the existing approaches.

Method	Employs spatial relationship	Employs temporal information	Employs place category distribution	Requires training	Requires external data
Graaff (2016) [43], Noureddine (2020) [44]	**Yes**	No	No	**No**	**No**
Gong (2016) [6]	**Yes**	**Yes**	No	**No**	Yes
Lv (2016) [7]	**Yes**	**Yes**	**Yes**	Yes	**No**
Zhang (2018) [42]	**Yes**	**Yes**	No	Yes	**No**
Zhang (2019) [45]	**Yes**	No	No	**No**	**No**
Bermingham (2019) [11]	**Yes**	No	**Yes**	**No**	**No**
Gao (2020) [8]	**Yes**	**Yes**	No	**No**	Yes
Our method	**Yes**	**Yes**	**Yes**	**No**	**No**

time probability for different place types from the activity log [6]. In addition, Gao adopted review data to determine the visiting time and duration probabilities of different place types [8]. The aforementioned methods either ignore the temporal information or rely on external data. Therefore, a probability model that fully makes use of spatiotemporal characteristics without using any external data is proposed in this paper. Table 1 summarizes the differences between our method and the existing approaches.

3. Methodology

3.1. Overview of the framework

The proposed semantic place annotation framework is illustrated in Fig. 1, and includes the trajectory collection, stop and candidate place extraction, spatiotemporal probability calculation of candidate places, and semantic annotation of visited places. The Unsupervised Place Annotation with Prior Probability (UPAPP) method is specifically developed in this work. First, the stops are extracted to segment a trajectory, and the relevant attributes of the stops are then calculated. Following this, a spatiotemporal probability model for the candidate places is created, which is decomposed into the spatial, duration and visiting time probabilities. Two different geospatial data are integrated to establish a spatial probability calculation method. For the duration and visiting time probabilities, we propose an unsupervised prior probability statistical method that does not require any supplementary data. The prior probabilities of different place types are extracted based on the potential visits of the trajectory. Once the spatiotemporal probabilities of the candidate places are obtained, the probability of visiting a place is calculated by the combination with the place type importance for the subsequent semantic annotation.

In the following, we provide some key definitions that are used throughout this paper.

Definition 1. Stop $SP = (x, y, t_{start}, dur)$ denotes a stop episode and is a list of continuous spatiotemporal trajectory points, where (x, y) is the coordinate of the stop center, which is calculated by the average coordinate of all spatiotemporal points in that stop; t_{start} represents the start time or visiting time of the stop; and dur indicates the duration calculated by the time difference between the end and the start of the stop.

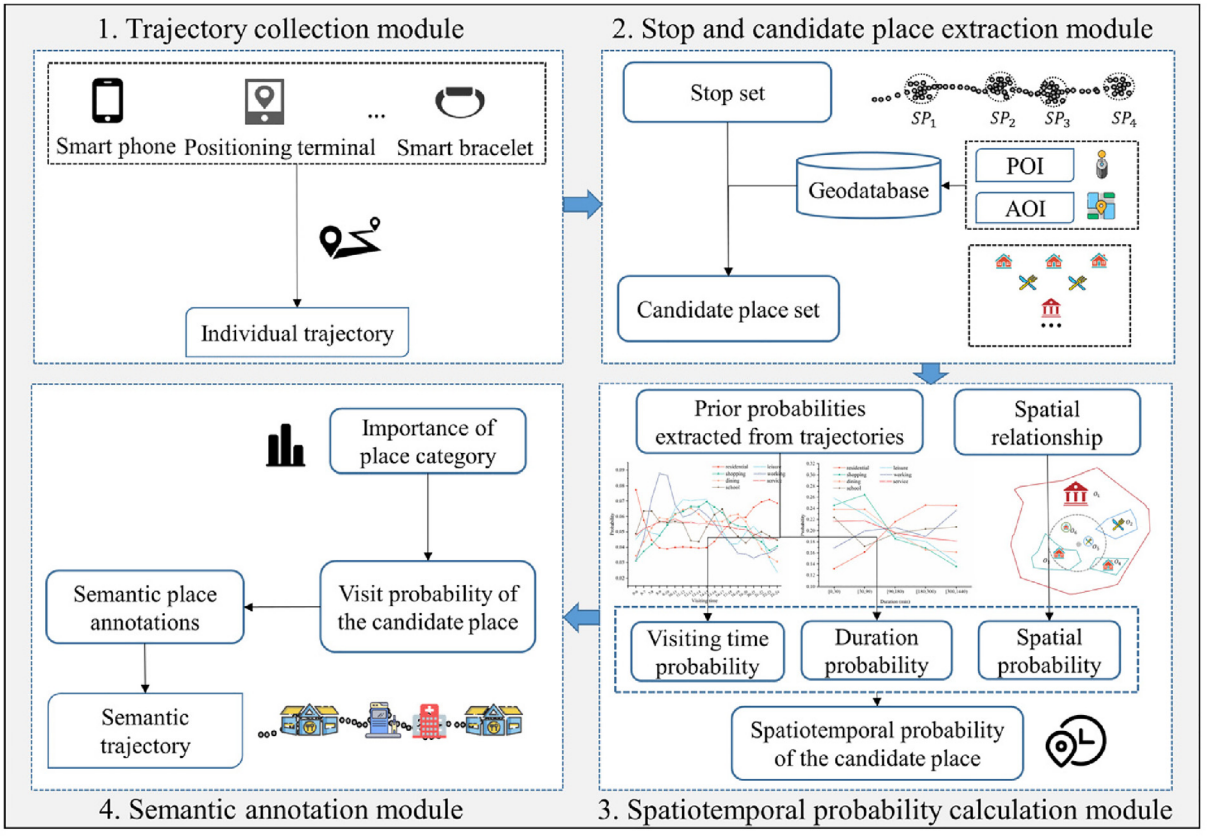

Fig. 1. Flowchart of the proposed UPAPP method.

Definition 2. Stop region $R_{SP} = (x, y, r)$ indicates the spatial range covered by the circle whose radius represents the spatial uncertainty and noise present in a real trajectory stop. In addition, the stop center is chosen as the center of the circle, and the stop radius equals the maximum distance to the stop center from all spatiotemporal points in the stop.

Definition 3. Candidate place O of the corresponding stop is a place with a distance from the stop center that is less than the search radius, where the search radius is user-specified.

Definition 4. The spatiotemporal probability of a candidate place indicates the probability of visiting the place under the corresponding spatial location and time conditions and is only related to the spatiotemporal conditions and the place itself.

Definition 5. The visit probability of a candidate place represents the probability of the place being visited. This variable is calculated by combining the spatiotemporal probability of the place and the place type importance and is related to the place category distribution in the candidate place set.

3.2. Extraction of the stops and candidate places

Three types of stop places were extracted [46]. The first category represents stops whereby the individual stays in a specific area without a positioning signal loss. This stop type is typically clustered based on spatiotemporal attributes. The second category denotes stops whereby the individual stays in the same area, yet the signal is partly lost due to object occlusion. The third includes stops that belong to the start and end of the trajectory, where the individual generally leaves or arrives at a room with no signal (e.g., leaving/arriving at home).

A clustering-based algorithm was chosen to detect the first category of stops. This algorithm searches for high-density clusters with strong aggregation in the spatiotemporal dimension. More specifically, the neighborhood of a spatiotemporal point is defined as the longest continuous subsequence starting from this point. The spatial distance from the starting point is less than the specified distance threshold d_1 for all points in the sequence. The density is defined as the period of the sequence. Points with a density exceeding the period threshold t_1 were denoted as core points, otherwise they were marked as noise. These definitions were integrated into the DBSCAN algorithm to identify the stops in the trajectory.

For the second stop category, the time interval and spatial distance of adjacent points were calculated to determine all point pairs whose time interval exceeds a specified time threshold t_2. The point pair is considered a stop if the distance between the point pair is less than a specified distance threshold d_2. The third category was extracted using the spatial distance between the end and starting points of the daily trajectory; if this distance is less than a specified distance threshold d_3, it is considered as a stop. Once the detection process was completed, all stops that were adjacent in both time and space were merged to ensure that the entire stop would not be divided into several smaller stops.

After that, the attributes of each stop were calculated (e.g., stop center, radius, start time, and duration), and the corresponding geospatial data (i.e., ROI and POI) was employed to search for candidate places around each stop. These two geospatial data sources were comprehensively integrated to obtain a complete geographic database and spatial understanding.

3.3. Calculating the spatiotemporal probabilities of candidate places

The spatiotemporal probability of a candidate place denotes the probability of a place being visited under the conditions of the corresponding spatial location and temporal attributes. The spatiotemporal probability of candidate place O_i corresponding to the stop point SP can be expressed as:

$$P(O_i|SP) = P(O_i|(x,y),t,dur), \tag{1}$$

where (x, y) is the coordinates of the stop center of SP; dur is the duration of SP; and t is the visiting (start) time of SP;

According to the Bayesian criterion, the probability can be calculated using *Eq.* (2).

$$P(O_i|(x,y),t,dur) = \frac{P((x,y),t,dur,O_i)}{P((x,y),t,dur)}, \tag{2}$$

where $P((x,y),t,dur,O_i)$ represents the joint distribution probability of stop SP and place O_i; and $P((x,y),t,dur)$ is not relevant to the place, for the same stop is a constant.

Based on the Bayesian criterion, $P((x,y),t,dur,O_i)$ is calculated as:

$$\begin{aligned} \mathrm{P}((x,y),t,dur,O_i) &= P(t|(x,y),dur,O_i)\cdot \mathrm{P}((x,y),dur,O_i) = P(t|(x,y),dur,O_i)\cdot P(dur|O_i,(x,y))\cdot P(O_i|(x,y)) \\ &\quad \cdot P((x,y)) \end{aligned} \tag{3}$$

Substituting Eq. (3) into Eq. (2) results in:

$$P(O_i|(x,y),t,dur) = \frac{P(t|(x,y),dur,O_i)\cdot P(dur|O_i,(x,y))\cdot \mathrm{P}(O_i|(x,y))\cdot P((x,y))}{\mathrm{P}((x,y),t,dur)}. \tag{4}$$

If we assume that (x, y), t, and dur are conditionally independent with respect to O_i, then we get:

$$P(O_i|(x,y),t,dur) = P(t|O_i)\cdot P(dur|O_i)\cdot P(O_i|(x,y))\cdot \frac{P((x,y))}{P((x,y),t,dur)}, \tag{5}$$

where $\frac{P((x,y))}{P((x,y),t,dur)}$ is a constant for the same stop. Therefore, if only the first three expressions (i.e., the visiting time, duration, and spatial probabilities, respectively) are considered, Eq. (5) can be expressed as$P(t|O_i)\cdot P(dur|O_i)\cdot P(O_i|(x,y))$. These three expressions are calculated in turn. The above three probabilities will simultaneously affect the annotation result of the visited place. As Fig. 2 shows, Place O_1 is most likely to be visited according to the spatial relationship only, while place O_5 is most likely to be visited under the corresponding temporal conditions.

3.3.1. Calculation of the spatial probability

The spatial probability of a candidate place is calculated based on the corresponding spatial distance and topological information. Here two geographic data types are considered, the POI and ROI.

The spatial probability of ROI employs the topological relationship between the candidate geographic object and the stop region, which can be divided into the following three categories: contain, intersection and disjoint [11]. If we denote the stop region of stop SP as R_{SP} and the geographic range of place O_i as R_{O_i}, then the relative spatial probability of O_i can be expressed as:

$$P_{relative}(O_i|(x,y)) = \begin{cases} 1,\ R_{SP}\ contain\ R_{O_i}\ or\ R_{O_i}\ contain\ R_{SP} \\ P_r + (1-P_r)*\frac{Area_{R_{O_i}\cap R_{SP}}}{Area_{R_{SP}}},\ R_{O_i}\ intersect\ R_{SP} \\ P_r * \frac{Searchradius - d_{R_{O_i},center_{SP}}}{Searchradius - r_{SP}},\ R_{O_i}\ disjoint\ R_{SP} \end{cases}, \tag{6}$$

where *contain* indicates that the topological relationship is contained if the stop region contains the place geometry (or vice-versa); *intersect* denotes that the topological relationship is intersecting but not contained; *disjoint* indicates that the topological relationship is separated; P_r is a user-specified parameter that represents the relative spatial probability when the place geometry just intersects the stop region; $Area_{R_{O_i}\cap R_{SP}}$ represents the intersection area of R_{O_i} and R_{SP}; $Area_{R_{SP}}$ is the area of the stop region of SP; $d_{R_{O_i},center_{SP}}$ represents the minimum distance from the stop center of SP to the geographic object O_i; $Searchradius$ is the user-specified radius used to search candidate places; and r_{SP} is the stop radius of SP.

For the spatial probability calculation of POIs, it is assumed that the influence of each POI on the surrounding areas obeys a two-dimensional Gaussian distribution, where the relative probability decays with distance [5]. To ensure that the relative probability of the POI is consistent with that of the ROI, the probability is constrained to be 1 when the distance equals 0 and P_r when the distance equals the stop radius. Thus, the relative probability of the POI is described as:

$$P_{relative}(O_i|(x,y)) = P((x,y)|O_i) = exp\left(-\frac{{d_{O_i,center_{SP}}}^2}{2\sigma^2}\right) s.t. exp\left(-\frac{{r_{SP}}^2}{2\sigma^2}\right) = P_r, \tag{7}$$

where $d_{O_i,center_{SP}}$ is the distance from the geographic object to the center of SP; and σ is the Gaussian distribution parameter that is calculated via the constraint conditions.

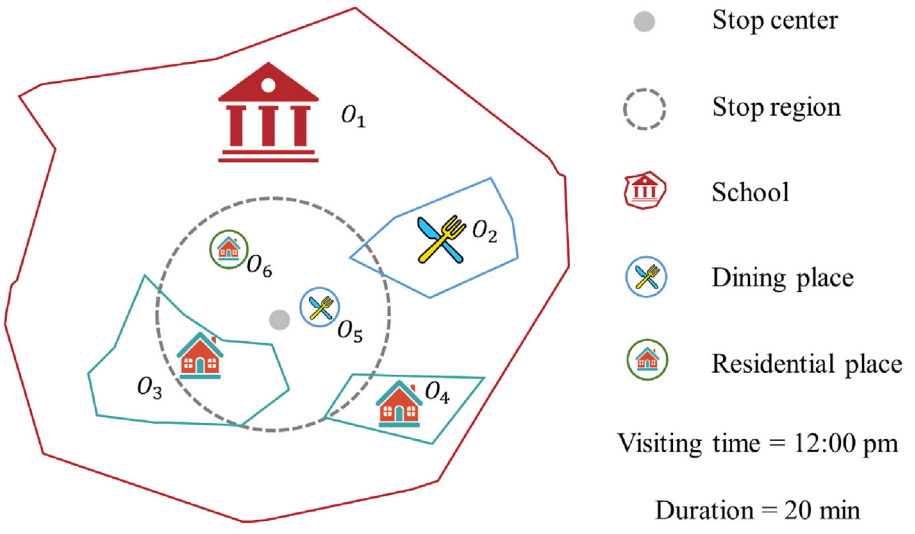

Fig. 2. An example of a stop and the corresponding candidate places.

Following the calculation of the relative spatial probability of all candidate places, the spatial probability is then determined via normalization. For candidate place O_i, the spatial probability is described as:

$$P(O_i|(x,y)) = \frac{P_{relative}(O_i|(x,y))}{\sum_i P_{relative}(O_i|(x,y))}, \tag{8}$$

where, $\sum_i P_{relative}(O_i|(x,y))$ is the sum of the relative spatial probability of all the candidate places in the corresponding stop.

3.3.2. Calculation of the duration probability

Visits to the same place categories follow the same duration and visiting time patterns [8]. Assuming that the place category corresponding to O_i is C_j, then the visiting time probability is $P(t|O_i) = P(t|C_j)$ and the duration probability is $P(dur|O_i) = P(dur|C_j)$. Previous work determines $P(t|C_j)$ and $P(dur|C_j)$ by counting labeled activity logs or extracting patterns from social media. However, this method relies on external data, which is not conducive to large-scale applications. In order to overcome the bottleneck of difficult-to-obtain annotated data, we propose a prior probability statistical method based on potential visits.

For the proposed method, a stop corresponds to an actual visit, and the places in the stop region are regarded as potential visits. These potential visits are weighted according to their importance. The Term Frequency-Inverse Document Frequency (TF-IDF) weighting method was adopted to calculate the importance weights of the potential visits [7]. If a stop region contains many places of the same type, and there are fewer places of this type in the whole area (e.g., the entire city), the importance of this type of place is high. The TF-IDF weighting method can effectively avoid the weight explosion and reflect the importance of the place category in the surrounding area. If the importance of category C_j places in the stop region is denoted as $I_{C_j}^{SP}$, then category C_j places are potentially visited $I_{C_j}^{SP}$*1 times in the corresponding stop.

When counting the duration prior probability, the duration is divided into multiple intervals $[dur_1, dur_2 \cdots dur_m \cdots]$. For stop SP with duration $dur \in dur_m$, the number of potential visits of stop SP to category C_j places is calculated as follows:

$$Visit_{C_j,dur_m}^{SP} = I_{C_j}^{SP} * 1 = \left(\frac{Count_{C_j}^{SP}}{\sum_j Count_{C_j}^{SP}} * log \frac{\sum_j Count_{C_j}}{Count_{C_j}} \right) * 1, \tag{9}$$

where $count_{C_j}^{SP}$ is the number of category C_j places in the SP stop region; $\sum_j Count_{C_j}^{SP}$ is the sum of the number of all places in the SP stop region; $Count_{C_j}$ is the number of category C_j places in the entire area; and $\sum_j Count_{C_j}$ is the sum of the number of all places in the entire area.

The above method is employed to calculate the number of potential visits in the entire trajectory dataset. The average number of potential visits of all stops with duration $dur \in dur_m$in the trajectory dataset to category C_j places is then determined as:

$$Visit_{C_j,dur_m} = \frac{1}{N_{dur_m}} \sum_{i=1}^{N_{dur_m}} Visit_{C_j,dur_m}^{SP_i}, \tag{10}$$

where $Visit_{C_j,dur_m}^{SP_i}$ is the number of potential visits to type C_j places in SP_i (duration dur$\in dur_m$); and N_{dur_m} is the number of stops with duration $dur \in dur_m$.

After separately counting the average number of potential visits to different place types belonging to different duration groups, the probability of visiting a certain place category when the duration belongs to a specified interval can be calculated. For example, the prior probability that the duration belongs to dur_m when visiting a category C_j place is given as:

$$P(dur \in dur_m|C_j) = \frac{Visit_{C_j,dur_m}}{\sum_m Visit_{C_j,dur_m}} = \frac{\frac{1}{N_{dur_m}} \sum_{i=1}^{N_{dur_m}} Visit_{C_j,dur_m}^{SP_i}}{\sum_m \frac{1}{N_{dur_m}} \sum_{i=1}^{N_{dur_m}} Visit_{C_j,dur_m}^{SP_i}}, \tag{11}$$

where $Visit_{C_j,dur_m}$ is the average number of potential visits of all stops with duration $dur \in dur_m$to category C_j places; and $\sum_m Visit_{C_j,dur_m}$ is the sum of the average number of potential visits across different duration intervals to category C_j places. An example of calculating the prior duration probability is presented in Appendix A.

Assuming that the candidate place O_i corresponds to place category C_j and the duration of the corresponding stop belongs to dur_m, then the duration probability of O_i is $P(dur|O_i) = P(dur|C_j) = P(dur \in dur_m|C_j)$.

3.3.3. Calculation of the visiting time probability

Similarly, the visiting time is divided into multiple intervals $[t_1, t_2 \cdots t_k \cdots]$ and calculated with Eq. (9), Eq. (10) and Eq. (11). The prior probabilities that the visiting time belongs to different intervals when visiting a certain place type are thus determined. For example, for a category C_j place, the prior probability of visiting time $t \in t_k$is calculated as follows:

$$P(t \in t_k|C_j) = \frac{Visit_{C_j,t_k}}{\sum_k Visit_{C_j,t_k}} = \frac{\frac{1}{N_{t_k}} \sum_{i=1}^{N_{t_k}} Visit_{C_j,t_k}^{SP_i}}{\sum_k \frac{1}{N_{t_k}} \sum_{i=1}^{N_{t_k}} Visit_{C_j,t_k}^{SP_i}}, \tag{12}$$

where $Visit_{C_j,t_k}$ represents the average number of potential visits of all stops with visiting time $t \in t_k$ to category C_j places; $\sum_k Visit_{C_j,t_k}$ is the sum of the average number of potential visits of stops to category C_j places with the visiting time belonging to different intervals; and $Visit_{C_j,t_k}^{SP_i}$ is the number of potential visits to type C_j places in SP_i with visiting time $t \in t_k$, which is the same as Eq. (9).

If candidate place O_i corresponds to place category C_j and the visiting time of the corresponding stop belongs to t_k, then the visiting time probability of O_i is $P(t|O_i) = P(t|C_j) = P(t \in t_k|C_j)$.

Finally, the spatiotemporal probability of the candidate place O_i is calculated as $P(t|O_i) \cdot P(dur|O_i) \cdot P(O_i|(x,y))$.

3.4. Annotation of visited places

In addition to the spatiotemporal attributes, the distribution of place categories also has an impact on the choice of visited places. For example, the probability of individuals visiting restaurants in areas containing many restaurants is higher. The probability of visiting the candidate place is essentially equal to the spatiotemporal probability of the place multiplied by the importance of the category to which the place belongs. If the candidate place set of SP is denoted as $O_{candidate} = (O_1, O_2 \cdots O_N)$, then for category C_j of place O_i, the importance of place O_i in $O_{candidate}$ is calculated as follows:

$$I_{O_i}^{O_{candidate}} = I_{C_j}^{O_{candidate}} = \frac{Count_{C_j}^{O_{candidate}}}{\sum_j Count_{C_j}^{O_{candidate}}} * log \frac{\sum_j Count_{C_j}}{Count_{C_j}}, \tag{13}$$

where $I_{O_i}^{O_{candidate}}$ indicates the importance of O_i in candidate place set $O_{candidate}$; C_j is the place category of O_i; $I_{C_j}^{O_{candidate}}$ indicates the importance of place category C_j; $Count_{C_j}^{O_{candidate}}$ is the number of candidate places of category C_j in $O_{candidate}$; and $Count_{C_j}$ is the number of places of category C_j in the entire area. The probability that O_i is visited in the corresponding stop is expressed as:

$$P(O_i) = P(O_i|(x,y), t, dur) * NI_{O_i}^{O_{candidate}}, \tag{14}$$

where $NI_{O_i}^{O_{candidate}}$ is the post-normalization importance of O_i.

For each stop, the candidate place with the highest visit probability is marked as the visited place.

4. Experiments and results

In order to evaluate the performance of our approach, we conducted comprehensive experiments using two real GPS trajectory datasets (i.e., the Shangdi-Qinghe dataset and the Geolife dataset) and various evaluation metrics. All the algorithms were implemented in Python and run on computers with Intel(R) Xeon(R) Silver 4116 CPU (2.10 GHz) and 128 GB memory.

4.1. Datasets

4.1.1. Trajectory datasets

(1) Shangdi-Qinghe trajectory datasets

The Shangdi-Qinghe trajectory dataset came from a daily behavior project conducted on the residents of Beijing City, China, from October to December 2012. This seven-day period project involved a total of 709 participants living or working in the Shangdi-Qinghe area, a subdistrict in Beijing [47,48]. GPS trackers equipped by volunteers recorded their locations every 30 s with a spatial positioning accuracy of 15 m. As the GPS trackers collected private locations, a privacy protection contract was signed by each participant. A total of 2,955,049 spatiotemporal points were collected, which is a sufficient representation of the target area.

While collecting the trajectory data, volunteers recorded their travel destination categories and activities on a daily basis to form activity logs. The collected information includes the ID, date, activity start time, activity end time, activity type, and place category. Place categories were grouped into residential areas, working areas, service venues, dining venues, schools, leisure venues, shopping venues, and others. The place category recorded in the activity log is used as the truth data to evaluate the accuracy of the UPAPP method in our experiment.

(2) Geolife trajectory datasets

The Geolife trajectory datasets were collected during the Microsoft Research Asia Geolife project period by 182 users from April 2007 to August 2012 [49]. This dataset contains 17,621 trajectories with a total duration of 50,176 h, recording a broad range of users' outdoor movements, including life routines and some entertainment and sports activities. The majority of the data was collected in Beijing, China, and participants were generally college students and Microsoft employees.

Noise is present in the original trajectory data due to signal blockages, and thus pre-processing is required. Following drifting, the data points considered as noise deviate from the original position in the route, producing abnormal speeds and included angles. The included angle refers to the angle formed by the connection between the central points and two points located at the front and back. We calculated the speed and included angle of each point, and removed those points with a speed exceeding 180 km/h or an angle less than 30° to eliminate the noise in the trajectory data.

4.1.2. Geographic data

The POI and ROI datasets were obtained from AutoNavi (the most famous navigation service company in China) and OSM (Open Street Map) data, respectively, in 2014. These geographic data were divided into the corresponding seven place categories in the activity logs, and the rules are described as follows.

The published POI data already includes classification information, and we further divided it into the following seven categories with corresponding keywords: 1) residential area: "community", "residential"; 2) working area: "company", "office building", "government"; 3) service place: "life services", "medical care", "finance", "car"; 4) dining place: "dining"; 5) school: "school"; 6) leisure place: "leisure"; and 7) shopping place: "shopping".

The elements of the OSM data contain a tag expressed in the form of key-value and were also divided into the following seven categories: 1) residential area: land use = residential; 2) working area: office = *; 3) dining place: amenity = restaurant; 4) school: amenity = college, amenity = university, amenity = school, amenity = kindergarten 5) leisure place: leisure = *; 6) shopping venue: shop = *; 7) service venue: amenity = *. Places belonging to other categories were removed out.

4.2. Experimental setup

4.2.1. Compared methods

To the best of our knowledge, the paper presents the first attempt at extracting the temporal probability from the trajectory itself. However, some existing unsupervised semantic annotation methods can be extended for place annotation. To verify the effectiveness of the proposed method, we compared the proposed UPAPP with other approaches by employing the following four methods: 1) spatial probability only; 2) spatiotemporal probability; 3) existing sequence model; and 4) the proposed method combined with a sequence model. The sequence model refers to using the order (i.e., sequence) of the visited places to establish a sequence model. Spatial and sequential characteristics have recently been employed to obtain a sequence model denoted as Spatio-Temporal Trajectories to Semantic Place-Matching Patterns (STOSEM) based on the HMM [11]. In this approach, the state transition probabilities were learned based on the potential visits.

The sequence models used for the comparison are detailed as follows: the observable sequence is the sequence of stop episodes; the hidden states correspond to candidate places associated with each stop episode that the individual may have stopped at; and the emission probability employs the probability of places visited described in Section 3.4. The state transition probability describes the probability of moving from one place category to another. Based on the potential visits, we learned all the probabilities in the trajectory to generate the state transition probability matrix. In particular, for adjacent stops SP_i and SP_{i+1}, there are a total of m category C_a places in the candidate places of SP_i, and a total of n category C_b places in the candidate places of SP_{i+1}. The potential visit calculation method introduced in Section 3.3 was applied to calculate the number of potential visits of SP_i (SP_{i+1}) to category C_a (C_b) places as $Visit_{C_a}^{SP_i}$ ($Visit_{C_b}^{SP_{i+1}}$). We assume that category C_b places are visited after type C_a places at a total of $Visit_{C_a}^{SP_1} * Visit_{C_b}^{SP_2}$ times, while the STOSEM method assumes this to occur $m*n$ times. All state transitions in the trajectory dataset were then counted to generate a state transition probability matrix. As a final step, the visited places were annotated via the Viterbi algorithm.

4.2.2. Evaluation metrics

To evaluate the effectiveness of the UPAPP method in semantic place annotation, we used the Shangdi-Qinghe trajectory dataset and the corresponding activity logs for a quantitative accuracy assessment. The accuracy for each place category was calculated using Eq. (15). Furthermore, the overall accuracy and average accuracy were also adopted to evaluate the effectiveness of semantic annotation.

$$acc_i = \frac{TP_i}{Total_i}, \tag{15}$$

where acc_i represents the accuracy of place category i; TP_i is the number of successful predictions of place category i; and $Total_i$ represents the total number of logged visits to place category i.

$$OA = \frac{\sum_i^N TP_i}{\sum_i^N Total_i}, \tag{16}$$

where OA is the overall accuracy, and N is the total number of place categories.

$$AA = \frac{\sum_i^N acc_i}{N}, \tag{17}$$

where AA is the average accuracy, and N is the total number of place categories.

4.3. Results and accuracy assessment

4.3.1. Extraction of stops

For the extraction of the first stop category, the time and space thresholds, t_1 and d_1, were set as 600 s and 100 m, respectively. Lost points belonging to the second type of stops are formed due to signal blockage, and consequently, the positioning accuracy of nearby points is relatively poor. For this category, we set the time and space thresholds t_2 and d_2 as 1200 s and 200 m, respectively. The spatial threshold d_3 of the third type of stops was also set as 200 m. When merging the above stops, the time and space thresholds were set as slightly lower than those of the first stop type, namely 90 m and 540 s, respectively. Furthermore, the search radius was selected as 200 m to match stops with candidate places.

In this experiment, 12,646 stops and 16,926 stops were extracted from the Shangdi-Qinghe dataset and the Geolife dataset, respectively. The distributions of all the detected stops from the two-trajectory dataset are presented in Fig. 3. The stops were observed to expand outwards from the center and were distributed across over half of the urban areas of Beijing city. The extracted stops in the Shangdi-Qinghe dataset were then matched with the activity log based on the corresponding temporal information. Among the extracted 12,646 stops, a total of 9,283 stops matched the activity log, while 1,345 stops were unsuccessfully matched. In addition, there were 1,848 stops that had no activity log on the corresponding date and 170 stops that were not surrounded by any candidate places.

4.3.2. Effectiveness of extracting the visiting time and duration probability

In order to extract the prior probability of the visiting time over 24 h, one day was divided into one group per hour. There were relatively few stops with the visiting time of 0–6:00 am, and thus they were merged into one group. As a result, the division resulted in 19 temporal groups. Due to the small temporal interval, a sliding window of three was employed to perform mean smoothing. Figs. 4(a) and 5(a) illustrate the prior probabilities of the visiting time for different place categories extracted from the Shangdi-Qinghe and Geolife datasets. Note that both datasets were collected in Beijing, and thus most of the extracted patterns were similar. Nevertheless, our method is able to efficiently capture differences in temporal regularity in different datasets. In general, the trends in the visiting time extracted from the trajectory data based on potential visits are essentially equal to residents' daily travel trends. The probability of visiting residential areas was higher in the morning and evening compared to visiting between 9:00 am and 4:00 pm. The probability of visiting working areas peaked at 8:00–10:00 am, with a second peak at 1:00–3:00 pm. Following 6:00 pm, the probabilities were lower, corresponding to the work patterns of individuals. The probability of visiting dining venues peaked at 11:00–1:00 pm and 6:00–7:00 pm, corresponding to lunch and dinner, respectively. Shopping was most probable in the afternoon, while the probability of visiting leisure places peaked at 2:00–4:00 pm and 8:00–10:00 pm. It is noteworthy that there were obvious differences in the trend of visiting schools in the two datasets. This can be attributed to the high participation of college students in the Geolife project, while the Shangdi-Qinghe dataset was completely based on information from individuals in employment. Furthermore, participants in the Geolife dataset had a preference for going out for leisure and shopping at night.

For the purpose of extracting the prior probability of the duration, the duration was divided into the following five groups: [0–30] min, [30–90] min, [90–180] min, [180–300] min, and [300–1440] min. The prior probabilities determined based on the method detailed in Section 3.3.2 are illustrated in Figs. 4(b) and 5(b). Likewise, the patterns extracted from

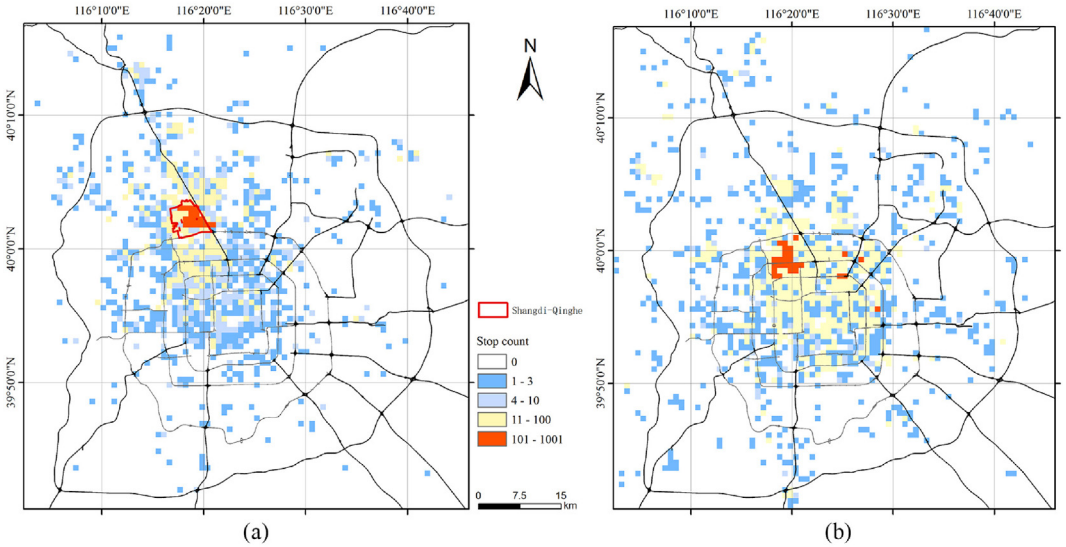

Fig. 3. Spatial distribution of extracted stops of the trajectory data. (a) the Shangdi-Qinghe dataset; (b) the Geolife Dataset.

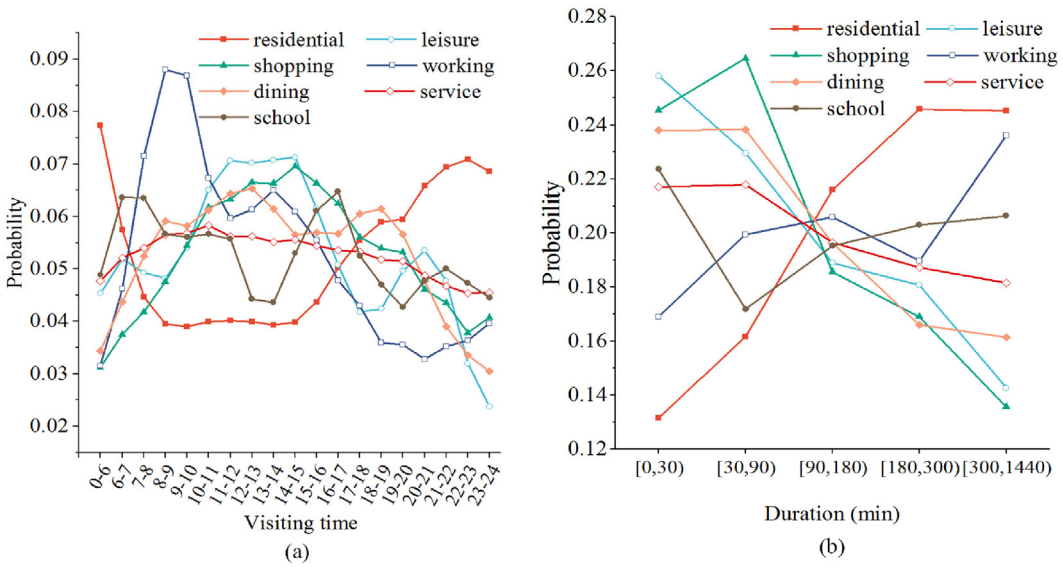

Fig. 4. Prior probabilities of visiting different place types extracted from the Shangdi-Qinghe trajectory dataset. (a) visiting time probabilities; (b) duration probabilities.

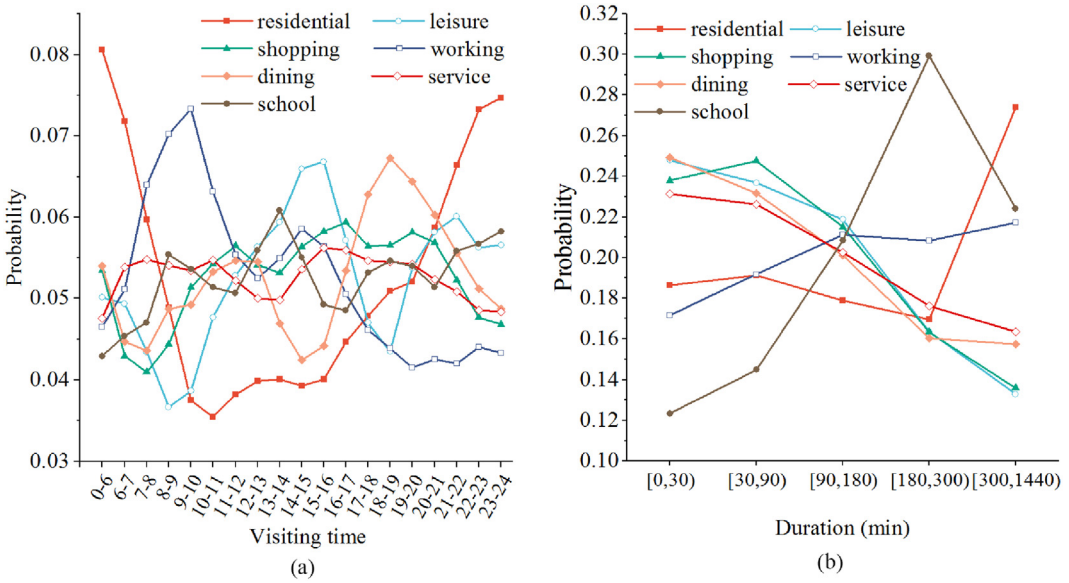

Fig. 5. Prior probabilities of visiting different place types extracted from the Geolife trajectory dataset. (a) visiting time probabilities; (b) duration probabilities.

the two datasets were consistent with the regularity of daily activities, while differences caused by different participants were also observed. The results demonstrate that when visiting places of leisure, dining, and services, the probability decreased with the duration. When visiting residential and working areas, longer durations were associated with a greater probability. When visiting shopping venues, the probability peaked at the 30–90 min duration group, and subsequently decreased as the duration increased. For school visits, durations of 0–30 min exhibited the highest probability in the Shangdi-Qinghe dataset. This is attributed to the minimal time required to pick up children. For the remaining duration groups, the longer the visiting time at schools, the higher the probability. In contrast, since the majority of participants lived at school, durations exceeding 180 min were more probable in the Geolife dataset.

The results indicate that UPAPP can reveal key temporal patterns from trajectory data without any external data. Moreover, the temporal patterns of visiting different place categories can also be quantified when activity logs are available. In order to verify the effectiveness of this method, we compared the extracted prior probabilities with the probabilities

retrieved from the activity log statistics of the Shangdi-Qinghe dataset (Fig. 6). In general, the temporal patterns for visiting different place types are essentially equal to the prior probabilities (Fig. 4). With respect to visiting time, visits to working, residential, dining, leisure, and shopping places exhibited similar probability peak positions. The patterns in duration determined from the activity log statistics are also generally consistent. However, there were also two differences observed between the two sets of results, described in the following. 1) Our results revealed relatively high probabilities for durations of 300–1440 min for workplace and school visits. This may be linked to the possible spatial proximity of schools and workplaces to lunch venues. At close distances, the morning and afternoon may be included in the same stop, thus extending the duration. 2) For leisure visits, our results indicated high probabilities for durations of 0–30 min. This was attributed to the multiple outdoor activities in leisure venues that were always moving, making it difficult to form a consistent stop. Overall, in the absence of activity logs, our method is effective in mining temporal patterns across places.

4.3.3. Effectiveness of semantic place annotation

Once the spatiotemporal probabilities of candidate places were calculated, the places visited in the stop episodes were annotated with the category importance and spatiotemporal probability. In this set of experiments, the search radius was set to 200 m and P_r was set to 0.5 for all the methods otherwise specified.

To determine the constancy between the actual and annotated place categories, the semantic annotation accuracy of the 9,283 stops in the Shangdi-Qinghe dataset that successfully matched the activity log was evaluated. Table 2 shows that the UPAPP method achieves the highest overall and average accuracy compared with others. Annotating using the spatial probability-only method has the lowest accuracy. The use of visiting time probability, duration probability, and influences of the surrounding places contribute to improving the accuracy. The results reveal the ability of the proposed method to effectively annotate the visited places, with relatively high recognition accuracies for all the categories. Furthermore, the use of the ROI data significantly improved the annotation accuracy of leisure venues, schools and residential areas. These place types typically have a large geographic range, and thus integrating spatial topological characteristics can more accurately reflect the spatial characteristics of stop events. The results indicate that combining the two data types allows for the comprehensive use of the topological characteristics and category distribution, hence effectively reflecting the spatial characteristics. Furthermore, in the absence of labeled data for training, sequence characteristics may not improve the model accuracy but rather reduce efficiency.

Since the Geolife dataset lacks labeled data for validation, we conducted a qualitative analysis of the annotation results. In this dataset, the most frequently-visited schools were Tsinghua University, Beihang University, Renmin University of China, and Peking University. Workplaces that appeared most frequently were the Yingdu building, the Chinese Academy of Sciences, the Sigma building (used by Microsoft Research Asia), and the Shenchang building. All the above places are located in the red areas in Fig. 3(b). This is consistent with the demographic statistics of the dataset [11,49]. Leisure places such as the Old Summer Palace, the Summer Palace, and the Olympic Forest Park near the aforementioned places were also frequently visited. Although a quantitative evaluation is lacking, the above analysis results also demonstrate the effectiveness of UPAPP.

In the second set of experiments, we examined the impact of varying parameters on the performance of the methods. First, we validated the effectiveness of the TF-IDF weighting algorithm in UPAPP. In our method, the TF-IDF weighting method is used to measure the importance of a place category in the surrounding area. We also compared the weighting

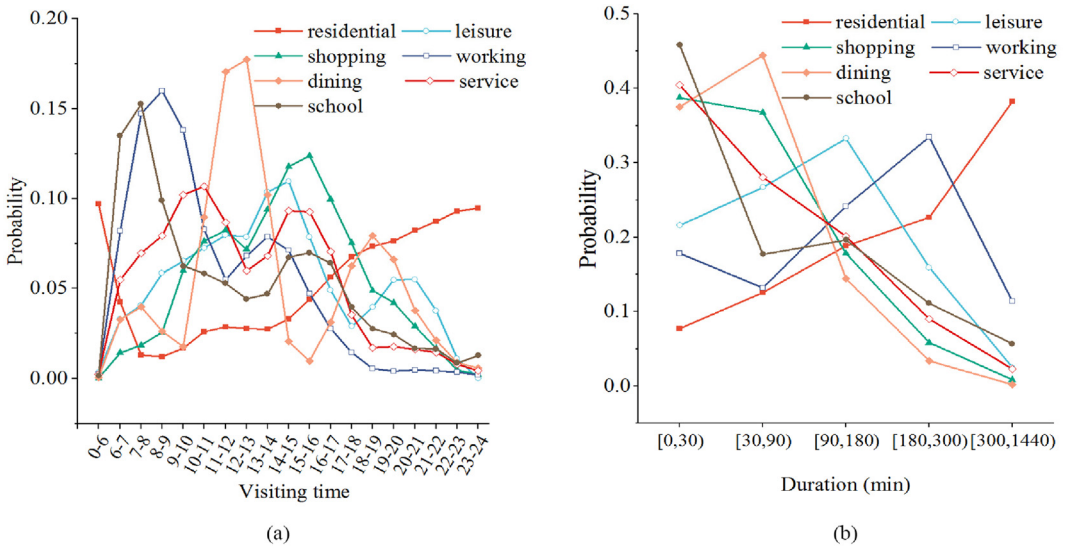

Fig. 6. Temporal probabilities for different place categories extracted from activity logs.

Table 2
Performance of semantic place annotation in the Shangdi-Qinghe dataset.

Performance metric	UPAPP	UPAPP combined with HMM	STOSEM	Spatial-only	Spatio-temporal	UPAPP with POI only
Dining place accuracy	**0.697**	0.574	0.563	0.406	0.480	0.691
School accuracy	**0.726**	0.567	0.375	0.713	0.632	0.449
Residential place accuracy	0.714	**0.788**	0.670	0.619	0.660	0.599
Leisure place accuracy	**0.701**	0.581	0.420	0.660	0.672	0.201
Working place accuracy	**0.682**	0.614	0.642	0.314	0.637	0.665
Shopping accuracy	0.819	0.717	**0.869**	0.437	0.633	0.759
Service place accuracy	0.700	0.672	**0.715**	0.529	0.399	0.650
Overall accuracy	**0.712**	0.706	0.652	0.529	0.623	0.616
Average accuracy	**0.720**	0.645	0.608	0.526	0.588	0.573

method using proportional, square root with using the logarithmic function. When the proportional method was directly used, the overall and average accuracy were 0.663 and 0.693, respectively. When the square root weighting method was used, the overall and average accuracy were 0.704 and 0.720, respectively. These results indicate that the TF-IDF weighting method is more powerful. Using the logarithmic function can effectively avoid the weight explosion and make the weight (i.e., IDF) of a certain category close to 0 when there are massive places of this category in the entire region.

We then compared the UPAPP with other methods to vary the search radius and spatial probability parameter P_r. As shown in Fig. 7, UPAPP outperforms the other methods under different parameters, which illustrates the stability and robustness of our method. Fig. 7(a) reveals that an inappropriate search radius affects the annotation accuracy. A small search radius may miss the correct place, while a large search radius will affect the distribution of surrounding places. Therefore, we recommend setting the search radius at 150–250 m empirically. Fig. 7(b) indicates that the methods are not sensitive to P_r for $P_r < 0.6$. When P_r is set too large, the spatial probability of places outside the stop radius will be relatively large, resulting in mismatching. Thus, we recommend setting P_r at 0.2–0.5.

5. Discussion

5.1. Practical contribution

To solve the dependence on annotated and external data in calculating temporal probability, we counted the visiting time and duration probability based on potential visits. Consequently, an unsupervised approach for semantic place annotation was proposed by combining spatial information, temporal information, and place category importance. To the best of the authors' knowledge, this is the first attempt at using temporal information without any external data. In other words, our method can be applied in different trajectory datasets that vary in regions and participants.

To further validate the effectiveness of our method, we annotated the visited places using the temporal probabilities derived from the activity logs. The overall accuracy of using probabilities determined by the log-based statistics is 0.725, and is slightly higher than our proposed method. This further demonstrates that the prior probabilities determined in our work can effectively reflect the temporal patterns of visits to multiple place types. Thus, without external data, our method

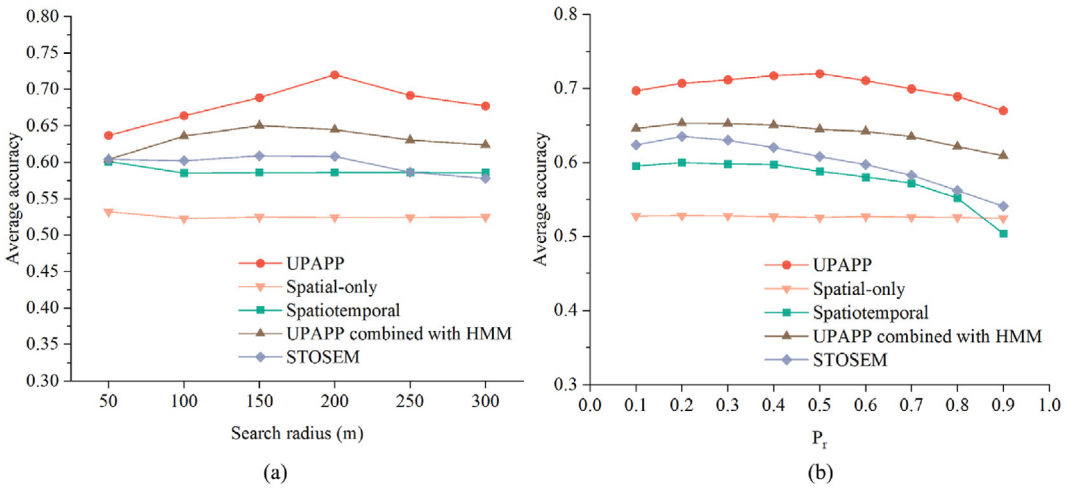

Fig. 7. Average accuracy of semantic place annotation under varying parameters on the Shangdi-Qinghe dataset. (a) search radius; (b) spatial probability parameter (P_r).

can make full use of spatiotemporal characteristics. This will greatly increase the utility of trajectory data for large-scale applications.

5.2. Limitations and future work

In Section 3.3, it is assumed that the location, visiting time and duration of a stop are conditionally independent at a given place. Considering that the visiting time and duration are both temporal attributes, this assumption may not necessarily be satisfied. Thus, we discuss the rationality of this assumption. Assuming that (x, y) and t are conditionally independent events with respect to O_i, while (x, y) and dur are conditionally independent with respect to O_i, the probability in Eq. (4) can be expressed as:

$$P(\mathrm{O_i}|(\mathrm{x,y}),\mathrm{t},\mathrm{dur}) = \frac{P\left(t|\mathrm{O_i},\mathrm{dur}\right)\cdot P(\mathrm{dur}|\mathrm{O_i})\cdot P(\mathrm{O_i}|(\mathrm{x,y}))\cdot P((x,y))}{\mathrm{P}((\mathrm{x,y}),\mathrm{t},\mathrm{dur})}. \tag{18}$$

As with Eq. (12), for the visited place type C_j and duration $dur \in dur_m$,the probability of visiting time $t \in t_k$can be calculated as follows:

$$P(t \in t_k|C_j, dur_m) = \frac{Visit_{C_j,dur_m,t_k}}{\sum_k Visit_{C_j,dur_m,t_k}} = \frac{\frac{1}{N_{dur_m,t_k}}\sum_{i=1}^{N_{dur_m,t_k}} Visit^{SP_i}_{C_j,dur_m,t_k}}{\sum_k \frac{1}{N_{dur_m,t_k}}\sum_{i=1}^{N_{dur_m,t_k}} Visit^{SP_i}_{C_j,dur_m,t_k}}, \tag{19}$$

where $Visit_{C_j,dur_m,t_k}$ represents the average number of potential visits of all stops with duration $dur \in dur_m$and visiting time $t \in t_k$to category $C_j$places; $\sum_k Visit_{C_j,dur_m,t_k}$ is the sum of the average potential visits of stops with duration $dur \in dur_m$to category C_j places when the visiting time belongs to different intervals; and N_{dur_m,t_k} is the number of stops with visiting time $t \in t_k$and duration $dur \in dur_m$.

Assuming that conditional independence between the duration and visiting time with respect to O_i are not satisfied, an overall accuracy of 0.710 is achieved, which does not reach the level of our method. However, this may be attributed to an insufficient amount of data. For the case of unconditional independence, the probability must be calculated under the two conditions, which requires a larger dataset. Increasing the trajectory data size to verify the conditional independence is reserved for future work. In addition, our method may consider the individual's historical trajectory in further work if the trajectory data is collected over a long period of time. For example, the individual's home and workplace can be extracted using the analysis of their multi-day trajectory to improve the annotation accuracy.

6. Conclusions

This paper presented UPAPP, an unsupervised method for the semantic place annotation of trajectories without the requirement of supplementary data. Specifically, trajectory stops were initially retrieved from the trajectory data, and the candidate places corresponding to each stop were identified. In order to infer the visited places, a spatiotemporal probability model was created, and the Bayesian Criterion was employed to decompose it into three terms: spatial probability, duration probability, and visiting time probability. A probability calculation method that integrates both POI and ROI geographic data was developed to calculate the spatial probability, fully utilizing the characteristics of the two data types. For the formulation of the visiting time and duration probabilities, the TF-IDF weighting algorithm was adopted to calculate the potential visits to different place types. By counting all the potential visits in the trajectory dataset, the prior probabilities of the visiting time and duration when visiting different place categories were generated. Following this, the spatiotemporal probability of the place was combined with the place category importance to annotate the visited place. Semantic annotation experiments were conducted on two real trajectory datasets. Results in the Shangdi-Qinghe trajectory dataset collected by 709 volunteers indicated that the UPAPP method achieved an overall accuracy of 0.712 and average accuracy of 0.720, outperforming several other methods and annotating visited places better without relying on any other data. Moreover, the temporal patterns acquired by potential visit statistics were essentially equal to those determined directly from the activity log. Without using annotated data, our method has great potential in large-scale automated trajectory annotation.

CRediT authorship contribution statement

Junyi Cheng: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Xianfeng Zhang:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition. **Peng Luo:** Methodology, Formal analysis. **Jie Huang:** Data curation, Software. **Jianfeng Huang:** Formal analysis.

Data availability

We published the code of our method and the Geolife trajectory dataset at https://figshare.com/articles/dataset/Geolife_data_and_UPAPP_code/18857615.

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.

Acknowledgments

The authors would like to thank Professor Yanwei Chai from the School of Urban and Environmental Sciences, Peking University, for providing the trajectory data and activity logs. This work was supported by two grants from the National Natural Science Foundation of China (No. 42171327) and the Xinjiang Production and Construction Corps, China (No. 2017DB005).

Appendix A. An example of calculating the prior duration probability

To demonstrate the calculation details of the probabilities in our UPAPP method, we present an example (Table 3) in which the durations are divided into five intervals: [0,30), [30–90), [90–180), [180–300), and [300,1440) min, and a total of 10 stops were used to count the prior probabilities. The potential visits to different types of places in Table 3 were calculated by Eq. (9).

Taking dining place visits as an example, we calculated the average number of potential visits of all stops with different duration intervals in the trajectory dataset to dining places using Eq. (10):

$$Visit_{\text{dining},\text{dur}\in[0,30)} = \frac{1}{N_{\text{dur}\in[0,30)}}\sum_{i=1}^{N_{\text{dur}\in[0,30)}} Visit^{SP_i}_{dining,N_{\text{dur}\in[0,30)}} = \frac{(0.26+0.31+0.27)}{3} = \frac{0.84}{3},$$

$$Visit_{\text{dining},\text{dur}\in[30,90)} = \frac{0.17+0.32}{2} = \frac{0.49}{2},$$

$$Visit_{\text{dining},\text{dur}\in[90,180)} = \frac{0.12+0.15+0.16}{3} = \frac{0.43}{3},$$

$$Visit_{\text{dining},\text{dur}\in[180,300)} = \frac{0.11}{1},$$

$$Visit_{\text{dining},\text{dur}\in[300,1440)} = \frac{0.09}{1}.$$

Next, the prior probabilities that the duration belongs to different intervals when visiting dining places were calculated with Eq. (11):

$$P(dur\in[0,30)|dining) = \frac{Visit_{dining,dur\in[0,30)}}{\sum_m Visit_{dining,dur_m}} = \frac{\frac{0.84}{3}}{\frac{0.84}{3}+\frac{0.49}{2}+\frac{0.43}{3}+\frac{0.11}{1}+\frac{0.09}{1}} = 0.32,$$

$$P(dur\in[30,90)|dining) = \frac{\frac{0.49}{2}}{\frac{0.84}{3}+\frac{0.49}{2}+\frac{0.43}{3}+\frac{0.11}{1}+\frac{0.09}{1}} = 0.28,$$

$$P(dur\in[90,180)|dining) = \frac{\frac{0.43}{3}}{\frac{0.84}{3}+\frac{0.49}{2}+\frac{0.43}{3}+\frac{0.11}{1}+\frac{0.09}{1}} = 0.17,$$

Table 3
An example of duration probability calculation.

Stop ID	Duration interval	Potential visits to schools	Potential visits to shopping places	Potential visits to dining places
1	[0,30)	0.12	0.19	0.26
2	[0,30)	0.14	0.16	0.31
3	[300,1440)	0.31	0.15	0.09
4	[30,90)	0.15	0.23	0.17
5	[90,180)	0.25	0.32	0.12
6	[180,300)	0.28	0.21	0.11
7	[30,90)	0.13	0.24	0.32
8	[0,30)	0.15	0.18	0.27
9	[90,180)	0.22	0.24	0.15
10	[90,180)	0.12	0.17	0.16

Table 4
Probabilities of visiting durations for different place categories.

Category	P(dur ∈ [0, 30)|C)	P(dur ∈ [30, 90)|C)	P(dur ∈ [90, 180)|C)	P(dur ∈ [180, 300)|C)	P(dur ∈ [300, 1440)|C)
Dining place	0.32	0.28	0.17	0.13	0.10
School	0.13	0.13	0.18	0.26	0.29
Shopping place	0.17	0.23	0.24	0.21	0.15

$$P(dur \in [180, 300)|dining) = \frac{\frac{0.11}{1}}{\frac{0.84}{3} + \frac{0.49}{2} + \frac{0.43}{3} + \frac{0.11}{1} + \frac{0.09}{1}} = 0.13,$$

$$P(dur \in [300, 1440)|dining) = \frac{\frac{0.09}{1}}{\frac{0.57}{2} + \frac{0.49}{2} + \frac{0.43}{3} + \frac{0.23}{2} + \frac{0.09}{1}} = 0.10.$$

Likewise, the prior duration probabilities of different categories of places were calculated (Table 4).

References

[1] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory pattern mining, In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, San Jose, California, USA, 2007, pp. 330-339. doi:10.1145/1281192.1281230.
[2] H. Cao, N. Mamoulis, D.W. Cheung, Discovery of periodic patterns in spatiotemporal sequences, IEEE Trans. Knowledge Data Eng. 19 (2007) 453–467, https://doi.org/10.1109/TKDE.2007.1002.
[3] C. Gao, Z. Zhang, C. Huang, H. Yin, Q. Yang, J. Shao, Semantic trajectory representation and retrieval via hierarchical embedding, Inf. Sci. 538 (2020) 176–192, https://doi.org/10.1016/j.ins.2020.05.107.
[4] Q. Gao, F. Zhou, T. Zhong, G. Trajcevski, X. Yang, T. Li, Contextual spatio-temporal graph representation learning for reinforced human mobility mining, Inf. Sci. 606 (2022) 230–249, https://doi.org/10.1016/j.ins.2022.05.049.
[5] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, K. Aberer, Semantic trajectories: mobility data computation and annotation, ACM Trans. Intel. Syst. Technol. 4 (2013) 1–38, https://doi.org/10.1145/2483669.2483682.
[6] L. Gong, X. Liu, L. Wu, Y. Liu, Inferring trip purposes and uncovering travel patterns from taxi trajectory data, Cartography Geograph. Inf. Sci. 43 (2016) 103–114, https://doi.org/10.1080/15230406.2015.1014424.
[7] M. Lv, L. Chen, Z. Xu, Y. Li, G. Chen, The discovery of personally semantic places based on trajectory data mining, Neurocomputing 173 (2016) 1142–1153, https://doi.org/10.1016/j.neucom.2015.08.071.
[8] Y. Gao, L. Huang, J. Feng, X. Wang, Semantic trajectory segmentation based on change-point detection and ontology, Int. J. Geog. Inf. Sci. 34 (2020) 2361–2394, https://doi.org/10.1080/13658816.2020.1798966.
[9] C. Wan, Y. Zhu, J. Yu, Y. Shen, SMOPAT: Mining semantic mobility patterns from trajectories of private vehicles, Inf. Sci. 429 (2018) 12–25, https://doi.org/10.1016/j.ins.2017.10.043.
[10] R. Fileto, M. Krüger, N. Pelekis, Y. Theodoridis, C. Renso, Baquara: A holistic ontological framework for movement analysis using linked data, In: International conference on conceptual modeling, Springer, Berlin, Heidelberg, 2013, pp. 342-355. doi:10.1007/978-3-642-41924-9_28.
[11] L. Bermingham, I. Lee, Mining place-matching patterns from spatio-temporal trajectories using complex real-world places, Expert Syst. Appl. 122 (2019) 334–350, https://doi.org/10.1016/j.eswa.2019.01.027.
[12] W. Xu, S. Dong, N. Venkateswaran, Application of artificial intelligence in an unsupervised algorithm for trajectory segmentation based on multiple motion features, Wireless Commun. Mobile Comput. 2022 (2022) 1–11.
[13] M. Etemad, A. Soares, E. Etemad, J. Rose, L. Torgo, S. Matwin, SWS: an unsupervised trajectory segmentation algorithm based on change detection with interpolation kernels, GeoInformatica 25 (2020) 269–289, https://doi.org/10.1007/s10707-020-00408-9.
[14] M. Etemad, Z. Etemad, A. Soares, V. Bogorny, S. Matwin, L. Torgo, Wise sliding window segmentation: a classification-aided approach for trajectory segmentation, in: Advances in Artificial Intelligence, Springer International Publishing, Cham, 2020, pp. 208–219, https://doi.org/10.1007/978-3-030-47358-7_20.
[15] A. Soares Junior, V. Cesario Times, C. Renso, S. Matwin, L.A.F. Cabral, A semi-supervised approach for the semantic segmentation of trajectories, in: 19th IEEE International Conference on Mobile Data Management (MDM), IEEE, Aalborg, Denmark, 2018, pp. 145–154, https://doi.org/10.1109/MDM.2018.00031.
[16] B. Ertl, M. Schneider, C. Diekmann, J. Meyer, A. Streit, A Semi-supervised Approach for Trajectory Segmentation to Identify Different Moisture Processes in the Atmosphere, In: Computational Science – ICCS 2021, Springer International Publishing, Cham, 2021, pp. 264-277. doi:10.1007/978-3-030-77961-0_23.
[17] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, W.-Y. Ma, Mining user similarity based on location history, In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems, ACM, Irvine, California, 2008, pp. 1-10. doi:10.1145/1463434.1463477.
[18] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.-Y. Ma, Understanding mobility based on GPS data, In: Proceedings of the 10th international conference on Ubiquitous computing, ACM, Seoul, Korea, 2008, pp. 312-321. doi:10.1145/1409635.1409677.
[19] T. Bhattacharya, L. Kulik, J. Bailey, Extracting significant places from mobile user GPS trajectories: a bearing change based approach, In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, ACM, Redondo Beach, California, 2012, pp. 398-401. doi:10.1145/2424321.2424374.
[20] A. Soares Júnior, B.N. Moreno, V.C. Times, S. Matwin, L.d.A.F. Cabral, GRASP-UTS: an algorithm for unsupervised trajectory segmentation, Int. J. Geog. Inf. Sci.: IJGIS 29 (2015) 46–68, https://doi.org/10.1080/13658816.2014.938078.
[21] M. Etemad, A. Soares, A. Hoseyni, J. Rose, S. Matwin, A trajectory segmentation algorithm based on interpolation-based change detection strategies, In: EDBT/ICDT Workshops, International Conference on Extending Database Technology/International Conference on Database Theory, Lisbon, Portugal, 2019. doi:10.13140/RG.2.2.34157.03049.
[22] M.L. Damiani, F. Hachem, H. Issa, N. Ranc, P. Moorcroft, F. Cagnacci, Cluster-based trajectory segmentation with local noise, Data Min. Knowl. Disc. 32 (2018) 1017–1055, https://doi.org/10.1007/s10618-018-0561-2.
[23] A.T. Palma, V. Bogorny, B. Kuijpers, L.O. Alvares, A clustering-based approach for discovering interesting places in trajectories, In: Proceedings of the 2008 ACM symposium on Applied computing, ACM, Fortaleza, Ceara, Brazil, 2008, pp. 863-868. doi:10.1145/1363686.1363886.
[24] L.H. Tran, Q.V.H. Nguyen, N.H. Do, Z. Yan, Robust and hierarchical stop discovery in sparse and diverse trajectories, Technical report at EPFL, No. EPFL-REPORT-175473, 2011. Available online: http://infoscience.epfl.ch/record/175473 (accessed on 27 May 2022).
[25] L. Gong, H. Sato, T. Yamamoto, T. Miwa, T. Morikawa, Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines, J. Modern Transp. 23 (2015) 202–213, https://doi.org/10.1007/s40534-015-0079-x.

[26] S. Hwang, C. VanDeMark, N. Dhatt, S.V. Yalla, R.T. Crews, Segmenting human trajectory data by movement states while addressing signal loss and signal noise, Int. J. Geog. Inf. Sci. 32 (2018) 1391–1412, https://doi.org/10.1080/13658816.2018.1423685.
[27] X. Niu, S. Wang, C.Q. Wu, Y. Li, P. Wu, J. Zhu, On a clustering-based mining approach with labeled semantics for significant place discovery, Inf. Sci. 578 (2021) 37–63, https://doi.org/10.1016/j.ins.2021.07.050.
[28] T.P. Nogueira, H. Martin, Querying semantic trajectory episodes, In: Proceedings of the 4th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS 2015, Seattle, USA, 2015, pp. 23-30. doi:10.1145/2834126.2834136.
[29] T.P. Nogueira, R.B. Braga, C.T. de Oliveira, H. Martin, FrameSTEP: A framework for annotating semantic trajectories based on episodes, Expert Syst. Appl. 92 (2018) 533–545, https://doi.org/10.1016/j.eswa.2017.10.004.
[30] R.D.S. Mello, V. Bogorny, L.O. Alvares, L.H.Z. Santana, C.A. Ferrero, A.A. Frozza, G.A. Schreiner, C. Renso, A multiple aspect view on trajectories, Trans. GIS (2019), https://doi.org/10.1111/tgis.12526.
[31] B. Zhao, M. Liu, J. Han, G. Ji, X. Liu, F.a. Zhu, Efficient semantic enrichment process for spatiotemporal trajectories, Wireless Commun. Mobile Comput. 2021 (2021) 1–13.
[32] F.J. Moreno, A.F. Pineda, R. Fileto, V. Bogorny, SMOT+: Extending the SMOT algorithm for discovering stops in nested sites, Comput. Inf. 33 (2014) 327–342.
[33] L. Gong, R. Kanamori, T. Yamamoto, Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons, Travel Behav. Soc. 11 (2018) 131–140, https://doi.org/10.1016/j.tbs.2017.03.004.
[34] T. Feng, H.J. Timmermans, Detecting activity type from GPS traces using spatial and temporal information, Eur. J. Transp. Infrastruct. Res. 15 (2015) 662–674, https://doi.org/10.18757/ejtir.2015.15.4.3103.
[35] A. Yazdizadeh, Z. Patterson, B. Farooq, An automated approach from GPS traces to complete trip information, Int. J. Transp. Sci. Technol. 8 (2019) 82–100, https://doi.org/10.1016/j.ijtst.2018.08.003.
[36] Y. Cui, C. Meng, Q. He, J. Gao, Forecasting current and next trip purpose with social media data and Google Places, Transp. Res. Part C: Emerg. Technol. IEEE Trans. Big Data 97 (2018) 159–174, https://doi.org/10.1016/j.trc.2018.10.017.
[37] C. Meng, Y. Cui, Q. He, L. Su, J. Gao, Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, Boston, MA, USA, 2017, pp. 1319–1324, https://doi.org/10.1109/BigData.2017.8258062.
[38] P. Stopher, C. FitzGerald, J. Zhang, Search for a global positioning system device to measure person travel, Transp. Res. Part C: Emerg. Technol. 16 (2008) 350–369, https://doi.org/10.1016/j.trc.2007.10.002.
[39] W. Bohte, K. Maat, 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. 17 (2009) 285–297, https://doi.org/10.1016/j.trc.2008.11.004.
[40] G. Xiao, Z. Juan, C. Zhang, Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization, Transp. Res. Part C: Emerg. Technol. 71 (2016) 447–463, https://doi.org/10.1016/j.trc.2016.08.008.
[41] H. Martin, D. Bucher, E. Suel, P. Zhao, F. Perez-Cruz, M. Raubal, Graph convolutional neural networks for human activity purpose imputation, In: NIPS spatiotemporal workshop at the 32nd Annual conference on neural information processing systems (NIPS 2018), Montreal, Canada, 2018. doi:10.3929/ethz-b-000310251.
[42] Y. Zhang, H. Wei, X. Lin, F. Wu, Z. Li, K. Chen, Y. Wang, J. Xu, Context-Aware Location Annotation on Mobility Records Through User Grouping, In: Advances in Knowledge Discovery and Data Mining, Springer, Cham, 2018, pp. 583-596. doi:10.1007/978-3-319-93040-4_46.
[43] V. De Graaff, R.A. De By, M. De Keulen, Automated semantic trajectory annotation with indoor point-of-interest visits in urban areas, In: Proceedings of the ACM Symposium on Applied Computing, ACM, Pisa, Italy., 2016, pp. 552-559. doi:10.1145/2851613.2851709.
[44] H. Noureddine, C. Ray, C. Claramunt, Semantic trajectory modelling in indoor and outdoor spaces, In: 2020 21st IEEE International Conference on Mobile Data Management (MDM), IEEE, Versailles, France, 2020, pp. 131-136. doi:10.1109/MDM48529.2020.00035.
[45] D. Zhang, K. Lee, I. Lee, Semantic periodic pattern mining from spatio-temporal trajectories, Inf. Sci. 502 (2019) 164–189, https://doi.org/10.1016/j.ins.2019.06.035.
[46] Z. Fu, Z. Tian, Y. Xu, C. Qiao, A two-step clustering approach to extract locations from individual GPS trajectory data, ISPRS Int. J. Geo-Inf. 5 (2016) 166, https://doi.org/10.3390/ijgi5100166.
[47] Tana, M.-P. Kwan, Y. Chai, Chai, Urban form, car ownership and activity space in inner suburbs: a comparison between Beijing (China) and Chicago (United States), Urban Stud. 53 (9) (2016) 1784–1802.
[48] L. Wu, L. Yang, Z. Huang, Y. Wang, Y. Chai, X. Peng, Y. Liu, Inferring demographics from human trajectories and geographical context, Comput. Environ. Urban Syst. 77 (2019), https://doi.org/10.1016/j.compenvurbsys.2019.101368 101368.
[49] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from GPS trajectories, In: Proceedings of the 18th international conference on world wide web, ACM, Madrid, Spain, 2009, pp. 791-800. doi:10.1145/1526709.1526816.