Computers, Environment and Urban Systems 77 (2019) 101346

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data

Mingxiao Li[a,b,c], Song Gao[c], Feng Lu[a,d,e], Hengcai Zhang[a,*]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[b] University of the Chinese Academy of Sciences, Beijing 100049, China
[c] Geospatial Data Science Lab, Department of Geography, University of Wisconsin, Madison, WI 53706, USA
[d] Fujian Collaborative Innovation Center for Big Data Applications in Governments, Fuzhou 350003, China
[e] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

ARTICLE INFO

Keywords:
Trajectory reconstruction
Mobile phone data
Missing data patterns
Data partitioning
Machine learning

ABSTRACT

Understanding human mobility is significant in many fields, such as geography, transportation, and sociology. Due to the wide spatiotemporal coverage and low operational cost, mobile phone data have been recognized as a major resource for human mobility research. However, due to conflicts between the data sparsity problem of mobile phone data and the requirement of fine-scale solutions, trajectory reconstruction is of considerable importance. Although there have been initial studies on this problem, existing methods rarely consider the effect of similarities among individuals and the patterns of missing data. To address this issue, we propose a multi-criteria data partitioning trajectory reconstruction (MDP-TR) method for large-scale mobile phone data. In the proposed method, a multi-criteria data partitioning (MDP) technique is used to measure the similarity among individuals in near real-time and investigate the spatiotemporal patterns of missing data. With this technique, the trajectory reconstruction from mobile phone data is then conducted with machine learning models. We verified the method using a real mobile phone dataset in a large city. Results indicate that the MDP-TR method outperforms competing methods in both accuracy and robustness. We argue that the MDP-TR method can be effectively utilized for grasping highly dynamic human movement status and improving the spatiotemporal resolution of human mobility research.

1. Introduction

Human movement is a major dynamic in various spatial and temporal phenomena, such as urban commuting, goods transportation, the spread of infectious diseases, and automobile pollutant diffusion (Gao, 2015; González, Hidalgo, & Barabási, 2008; Kang, Ma, Tong, & Liu, 2012; Xu, Belyi, Bojic, & Ratti, 2018; Yao et al., 2018). Thus, studies on human mobility are crucial in numerous domains including urban planning, intelligent traffic management, and ecology (K. Liu, Gao, & Lu, 2019; Yue, Lan, Yeh, & Li, 2014). With the development of information and communications technologies (ICTs), positioning technologies, and the popularization of intelligent terminals, massive quantities of trajectory data are quickly accumulated (Cao et al., 2015; Li, Lu, Zhang, & Chen, 2018; Liu, Gong, Gong, & Liu, 2015; Zheng, 2015). The availability of such data brings new opportunities to uncover the human mobility and the complex interplay between the urban environment and human activity (Giannotti et al., 2011; Sui & Shaw, 2018; Wan et al., 2018).

Due to its wide spatiotemporal coverage and low operational cost, mobile phone data have been recognized as a major data source for many mobility related studies and applications (Gurumurthy & Kockelman, 2018; Shin et al., 2015; Xu et al., 2018; Pei et al., 2014). Call detail record data (CDR) is one of the most accessible and popular types of mobile phone data. It records human movement when a mobile phone interacts with a phone tower (Ahas, Aasa, Silm, & Tiru, 2010; Gao et al., 2017; Gao, Liu, Wang, & Ma, 2013; Yuan, Raubal, & Liu, 2012). To reduce the cost of data transmission and storage space imposed by the massive mobile communication systems, records are collected only at each occurrence of a phone communication activity (i.e., phone calls or text messages) (Deville et al., 2014; Song, Qu, Blumm, & Barabási, 2010). This inevitably results in the problem of trajectory discontinuity (Ranjan, Zang, Zhang, & Bolot, 2012; Zhao et al., 2016),

* Corresponding author at: State Key Lab of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.
E-mail address: zhanghc@lreis.ac.cn (H. Zhang).

https://doi.org/10.1016/j.compenvurbsys.2019.101346
Received 14 January 2019; Received in revised form 13 May 2019; Accepted 15 May 2019
0198-9715/ © 2019 Elsevier Ltd. All rights reserved.

which brings about inaccurate and even unreasonable analyses and inferences (Cheng & Lu, 2017; Deng, Fan, Liu, & Gong, 2016). Thus, the trajectory reconstruction of CDR data is a key preprocessing step in many applications.

Although previous studies have attempted to resolve the trajectory reconstruction problem (G. Chen, Viana, & Sarraute, 2017; Fan et al., 2016; Liu et al., 2018), two limitations remain. First, understanding the similarities between individuals can enhance reconstruction performance. However, since only a few methods have been developed for spatiotemporal proximity analysis of large-scale low-frequency movement data in near real-time (Yuan, Chen, Li, Shaw, & Lam, 2018), most methods use only individual trajectories or single trajectory segment to reconstruct the missing data. Reconstruction based on individual patterns tends to suffer from the data-hungry problem when an individual only has a small number of historical trajectories available. This indicates that meaningful movement patterns are unable to be mined and limit method versatility. Second, trajectory reconstruction methods are challenged by the variations in the temporal patterns of missing data. That is, the reconstructions of locations are quite different for missing data with different time spans. Moreover, missing data at different times may lead to different inferences with respect to various movement behaviors. Ignoring the differences in temporal patterns of missing data may reduce the precision and reliability of the trajectory reconstruction methods.

To overcome the limitations discussed above, we propose a new multi-criteria data partitioning trajectory reconstruction method, known as the MDP-TR method, to reconstruct missing points in large-scale mobile phone datasets. The contributions of this study can be summarized as follows:

(1) We propose an anchor-point-based clustering algorithm applied to the trajectory reconstruction area. This helps in understanding the similarities between large-scale low-frequency trajectory datasets in near real-time and helps lessen the data-hungry problem, which can significantly improve the performance and reliability of trajectory reconstruction.
(2) We introduce an incremental-sliding-window-based data construction algorithm. This can generate trajectory segments with different time spans and at different times, which allows our model to learn the relevant movement characteristics with different temporal patterns of missing data.
(3) We evaluate the proposed method using a large-scale continuous mobile phone dataset collected from 1,000,000 individuals over a period of 15 workdays, where the number of location points exceeded one billion. The results demonstrate significant improvements in reconstruction performance over competing solutions.

The remaining parts of this study are organized as follows. Section 2 provides a literature review and Section 3 states the problem. Section 4 describes the details of the proposed multi-criteria data partitioning trajectory reconstruction (MDP-TR) method. The experimental results and a performance analysis are presented in Section 5. The final section concludes this study and discusses future research.

2. Literature review

2.1. Trajectory reconstruction for mobile phone data

In the literature, previous studies have proposed several trajectory reconstruction methods. These related studies can be roughly divided into three categories: map-matching-based methods, interpolation-based methods, and pattern-learning-based methods. The basic assumption of map-matching-based methods is that individual movement behaviors follow the road networks. Thus, a sequence of trajectory points can be aligned to a sequence of road segments to form a complete path (Algizawy, Ogawa, & El-Mahdy, 2017; Chen & Bierlaire, 2015; Jagadeesh & Srikanthan, 2015). Jagadeesh and Srikanthan (2017) used a hidden Markov model to generate partial map-matched paths and then used a pre-training route choice model to identify the most likely path. Xiao, Wen, Markham, and Trigoni (2014) used contextual relationships between trajectory points as features of the CDR trajectories in a conditional random field model to reconstruct individual trajectories. However, the basic assumption that underlies the map-matching method is questionable. That is, individuals in urban space can travel by subway or on foot, which limits the performance of such methods. In addition, these methods mainly concern the spatial complement and seldom attempts to estimate the time required for an individual to reach the road segments. These factors may limit the applicability of such methods.

Interpolation-based methods mainly use spatial-temporal correlations among data to interpolate missing points. The main assumption when using this type of methods is that missing points can be approximated by a simple function (e.g., nearest-neighbor function, linear function, or Gaussian function). During interpolation, each trajectory point is assumed to be independent of the others and weights are calculated based on the distances and time spans between each missing point and its contextual points (Ficek & Kencl, 2012; Hoteit, Secci, Sobolevsky, Pujolle, & Ratti, 2013; Hoteit, Secci, Sobolevsky, Ratti, & Pujolle, 2014; Yu, Russell, Mulholland, & Huang, 2018). However, the movement patterns of individuals are complex. Especially when the time span between two given points is long (which is quite common in CDRs), the individual is likely to either have traveled to several locations or to have stayed in the same location. Thus, in such cases, reconstruction is ineffective.

With recent advances in artificial intelligence, pattern-learning-based reconstruction methods have become mainstream in this discipline. They can be used to improve the models by exploring the individual behavioral characteristics from historical trajectories and performing trajectory reconstruction using trained models, such as the Markov models, topic models, and neural network models (Chen et al., 2017; Fan et al., 2016; Liu et al., 2018). For example, Fan et al. (2016) proposed a collaborative filtering method for CDR reconstruction. They inferred individual movement patterns using a topic model and then estimated the location sequence with a hidden Markov model using the topic distributions. Liu et al. (2018) divided individuals into three types based on an indicator of the radius of gyration and used a back propagation neural network to reconstruct the respective positions of mobile phone users. Whereas, the accuracies of these methods were not satisfactory and required further improvement.

2.2. Anchor point detection

Anchor points represent the locations of key activities in people's daily lives, which can provide theoretical support for commuting behavior research and human mobility research (Hägerstraand, 1970; Kung, Greco, Sobolevsky, & Ratti, 2014; Miller, 1991; Palma, Bogorny, Kuijpers, & Alvares, 2008). Traditional techniques to obtain anchor points mainly rely on surveys (Cascetta, 1984) or modeling methods (Han, Hao, Wang, & Zhou, 2011; Yan, Han, Zhou, & Wang, 2011). However, problems, such as high cost, poor timeliness, and extended update periods, limit performance. With the development of ICT technology, massive trajectories provide two ways for anchor point: semantic rule-based and moving state derivation-based methods.

The basic idea of semantic rule-based methods is to refer to urban resident living habits and identify the locations where the resident stays the longest during working hours (e.g., 08:00 to 18:00) and rest hours (e.g., 22:00 to 06:00) as the workplace and residence, respectively (Calabrese, Diao, Di Lorenzo, Ferreira Jr, & Ratti, 2013; Kung et al., 2014). These methods exhibit significant performance in specific categories. However, these methods rely on prior knowledge and can only determine anchor points of limited categories, which restricts the applicability of such methods. The moving state derivation-based method

typically selects activity range, dwell time, trajectory point density or moving speed as a threshold parameter. When an individual stays within a certain range during a period of time (Capela et al., 2016) or the trajectory points have obvious agglomeration pattern in space (Huang, Cao, & Wang, 2014; Palma et al., 2008), the cluster of those points is identified as an anchor point. These methods can extract the activity sequence patterns from individual trajectories, which is crucial for individual behavior understanding and personalized recommendation. However, how to make the anchor points semantic is a key challenge.

2.3. Trajectory similarity measurement

When clustering trajectories, we need to calculate the similarity among trajectories. The traditional techniques to measure trajectory similarity mainly concerned on the spatial proximity. These methods measured the similarity using the distances between entire trajectories via algorithms such as Closet-Pair distance, Dynamic Time Wrapping (DTW) distance, and Longest Common Sub-Sequence (LCSS) distance (Chen, Özsu, & Oria, 2005; Chen, Shen, Zhou, Zheng, & Xie, 2010; Toohey & Duckham, 2015) or the distance between the trajectory segments (Lee, Han, & Whang, 2007; Mao, Zhong, Qi, Ping, & Li, 2017). However, the trajectories are with time attribute, and only considering the spatial similarity will lead to misidentification of the spatiotemporal neighbor relationship.

To the best of our knowledge, few methods focus on the spatiotemporal proximity analysis of trajectory data. Several studies used brute force approaches (Zheng, 2015), space-time separated approaches (Long & Nelson, 2013; Miller, 2005), or space-time buffering approaches (Yuan et al., 2018) for similarity measurement. However, these methods are typically computationally intensive, which are not applicable to datasets with millions of trajectories in near real-time. Therefore, new less time-consuming methods are required for spatiotemporal proximity analysis.

3. Problem statement

As discussed previously, the problem of interest is to reconstruct the points of the individual trajectories that are missing in a large-scale CDR dataset. Suppose that $Traj_{\text{full}} = \langle pt_1, pt_2, \ldots, pt_n \rangle$ represents the full trajectory of an individual and $Traj_{\text{record}} = \langle pt_1, pt_2, \ldots, pt_m \rangle$ represents the recorded trajectory of that same individual, where pt_i represents a trajectory point that includes latitude, longitude, and timestamp information (x_i, y_i, t_i); $\forall i \in [1, n]$, $t_i < t_{i+1}$. If there exists $\{pt_j \mid pt_j \in Traj_{\text{full}}$ and $pt_j \notin Traj_{\text{record}}\}$, it indicates that pt_j is a missing point in $Traj_{\text{record}}$. To solve this problem, the frequent movement patterns and the spatiotemporal characteristics of the historical trajectories and the most likely locations are estimated based on prior knowledge.

Understanding the similarity among individuals can enhance trajectory reconstruction performance. However, due to the large-scale characteristics and inconsistent recording times in mobile phone data, most methods only use individual trajectories or single trajectory segment to reconstruct missing data. Limited amounts of data lead to the data-hungry problem, which makes the model overfitting and limits the effectiveness of the method. Anchor points represent the locations of key activities in people's everyday lives (Hägerstraand, 1970; Miller, 2005; Xu et al., 2015). As individuals in urban space tend to spend most of their time in a few specific locations, the trajectory can be roughly represented as stops and moves through a set of interesting places (Long & Thill, 2015; Moreno, Pineda, Fileto, & Bogorny, 2014; Song et al., 2010). These places denote the locations of an individual's major activities, such as the home, workplace, and transfer stations, which significantly reflect the characteristics of that individual's movements. This provides an effective way to measure the similarity among large-scale low-frequency trajectories using anchor points. For example, the individuals with similar workplace and home locations are more likely to select the same path for their commute and leave home at similar times given their spatiotemporal coupling constraints. Therefore, individuals with similar movement patterns will be clustered into a group whose trajectories will be used to provide a more adequate dataset for the model training.

The temporal pattern of the missing data affects trajectory reconstruction results. For example, suppose that we want to reconstruct the trajectory between two locations that are 3 km apart. If the time span between the two records is approximately 5 min, the missing point is most likely to occur along the shortest path between the two locations. If the time span is approximately 5 h, there will be more distributions possible for the missing data. To consider these impacts on trajectory reconstruction, we define the temporal patterns of the missing data as $TM\_Pattern_{pt_j} = \{time_{pt_j}, timespan_{pt_j}\}$, where $time_{pt_j}$ represents the time of day of the missing data and $timespan_{pt_j}$ represents the time interval between the previous record and the next record of the missing data.

4. Methodology

4.1. Framework

The method developed in this study (MDP-TR) consists of two parts: the multi-criteria data partitioning (MDP) technique and the MDP-TR modeling. As shown in Fig. 1, the method begins with an anchor point-based clustering algorithm. With this algorithm, individuals are divided into a series of groups in which each group represents several individuals whose movement patterns are similar, which enhances reconstruction performance by solving the data-hungry problem. Then, an incremental-sliding-window-based data construction algorithm is

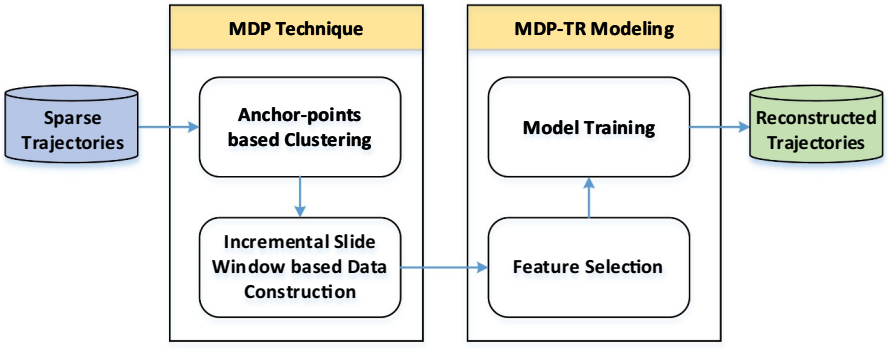

Fig. 1. The reconstruction process for the multi-criteria data partitioning trajectory reconstruction (MDP-TR) method.

proposed. This algorithm extracts the individual's trajectories in each group separately, and divides the trajectories into trajectory segments with varying time spans at different times to take into account different temporal patterns of the missing data. Finally, we select the features of trajectory segments in each group for model training and construct the MDP-TR model by importing the MDP technique into classic machine learning models. As a consequence of the previous operator, a series of MDP-TR models are trained to reconstruct the individual trajectories of each group.

4.2. Anchor-point-based clustering algorithm

Considering the complexity of individuals' movement patterns and the high computational cost for calculating the similarities between large-scale movement data, most methods only use one individual's historical trajectory or single trajectory to reconstruct the missing data. This can lead to data-hungry problems and limit method versatility. Since anchor points can represent the locations of key activities in people's daily lives, in this subsection, we propose an anchor-point-based clustering algorithm into the area of trajectory reconstruction to gain an understanding the similarities among massive populations at a low computational cost in near real-time.

4.2.1. Anchor point detection

Due to the challenge that an individual's CDR locations may switch among adjacent cell phone towers caused by signal strength variation, we define an anchor point as a set of adjacent cell phone towers where the individual spends a certain minimum amount of time (Huang et al., 2010; Xu, Shaw, Fang, & Yin, 2016). To detect the anchor points, we first calculate the number of records associated with each cell phone tower in an individual's trajectories. Then, the most frequently visited tower is selected as the core and all neighboring towers, within a distance threshold of β, are grouped as a candidate. After that, the next most frequently visited tower is selected, and the same process is iteratively performed until all the individual's records are processed. Finally, the visit frequency of each candidate is calculated, and each candidate whose visit frequency exceeds 20% of an individual's total visits recorded in CDRs is defined as an anchor point of the individual. The most frequently visited tower in each anchor point is defined as the representation location. An example of anchor point identification for an individual is shown in Fig. 2, where the green lines represent an individual's spatiotemporal trajectory (also known as the space-time path (Hägerstraand, 1970; Miller, 1991; Shaw, Yu, & Bombom, 2008)) in 3-D space, the blue circles represent the anchor points in the 2-D plane, the red circle represents a random point set that contains less frequently visited locations, and the representation location of the individual's anchor points are the locations of cell phone towers 1, 2, and 3. Notice that we used circular regions rather than points in this figure to acknowledge the location uncertainty around each cell tower.

It is worth noting that the distance threshold, β, is crucial when

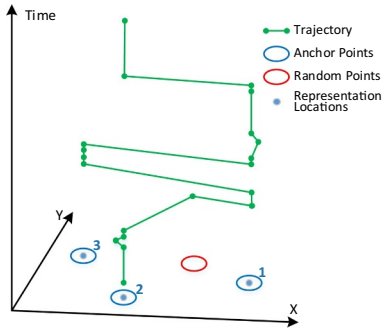

Fig. 2. An illustration of anchor point detection.

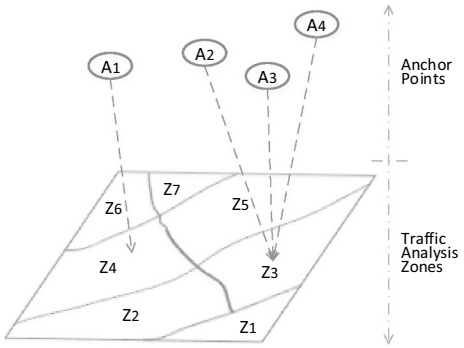

Fig. 3. The projections from the anchor points to the traffic analysis zones.

detecting anchor points. On the one hand, choosing a small value will not help lessen the cell phone tower switching problem caused by signal strength changes. On the other hand, a large value will fail to capture the individual's movements through space because most of the points would be aggregated into a single large cluster. In this study, considering that the average distance between nearest neighbor cell phone towers is approximately 0.24 km, we selected a 0.5 km as the distance threshold to achieve a compromise between the signal switching problem and the objective of movement identification in urban space. This threshold was also applied in the existing literature to cluster anchor points for the identification of individuals' home and job locations (Long & Thill, 2015).

4.2.2. Clustering with anchor points

As a consequence of the previous identification operator, each individual's trajectory is roughly represented as a set of anchor points. In this step, we aim to cluster individuals into a series of groups, where each group represents certain individuals who have similar movement patterns. First, we project the location of each individual's anchor points to traffic analysis zones (Martínez, Viegas, & Silva, 2009; Yuan et al., 2015) using its representation location, as shown in Fig. 3. Then, we calculate the similarity of each pair of individuals using the Jaccard index (Jaccard, 1912), which can be implemented with the following equation:

$$Sim_{(I_a,I_b)} = \frac{|Z_a \bigcap Z_b|}{|Z_a \bigcup Z_b|} \quad (1)$$

where $Sim_{(I_a,I_b)}$ denotes the similarity between the two individuals I_a and I_b, Z_a and Z_b denote the traffic analysis zone ID sets of the individuals, respectively, $|Z_a \bigcap Z_b|$ denotes the size of the intersection of the two sets, and $|Z_a \bigcup Z_b|$ denotes the size of the union of the two sets. After acquiring the similarity matrix for all individuals, we apply the agglomerative hierarchical clustering method (Gil-Garcia, Badia-Contelles, & Pons-Porrata, 2006) to generate the clusters. This method treats each individual as a cluster initially and then combines the clusters according to the nearest distance principle.

Note that there are three reasons that we project the anchor points into traffic analysis zones. First, we consider traffic analysis zones as basic spatial units in urban studies because people perform socio-economic activities inside these zones and this is the fundamental cause of human movement. Second, since each individual's anchor point locations are different, projecting the anchor points into traffic analysis zones allows us to convert the spatial distance calculation into a one-dimensional matching calculation. This significantly reduces computational costs, which allows the possibility of measuring the similarity among massive quantities of trajectories in near real-time. Last but not least, data with a spatial resolution that is finer than zonal data, such as census blocks, are not available in our study area.

4.3. Incremental-sliding-window-based data construction algorithm

As previously mentioned, the missing data points at different times and with varying time spans may lead to different inferences with respect to human movement behaviors. Hence, it is necessary to take this issue into account when reconstructing an individual's trajectory.

To address this problem, we propose an incremental-sliding-window-based data construction algorithm for the temporal pattern of the missing data. The core idea of the proposed algorithm is to design a series of sliding windows to partition the individuals' trajectories. The size of the sliding window is designed to gradually increment by one time step for each iteration to take into account the influence of the missing data with varying time spans. After dividing the trajectories into segments, the points within the sliding window are treated as label points in turn and a series of tuples are formed for each segment as $Seq_j = \{pt_s, pt_j, pt_e\}$ to consider the influence of data missing at different times. In the tuples, pt_s and pt_e represent the start point and end point of each segment and pt_j represents one of the missing points in the segment, $\forall j \in (s, e)$. Fig. 4 shows an illustration of the algorithm with one fixed start point, where the constructed segments have varying time spans and the label data occur at different times. With the incremental sliding window algorithm, we can construct the training data with different patterns of missing data, which helps us model the human movements more comprehensively and precisely.

Algorithm 1: Incremental-sliding-window-based data construction algorithm.

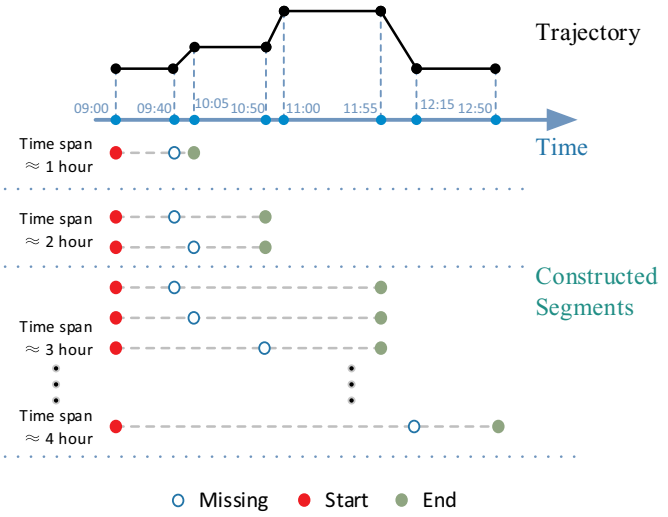

Fig. 4. An illustration of the incremental-sliding-window-based data construction algorithm.

Input: $Traj_k = \{pt_1, pt_2, \ldots, pt_m\}$, maximum time interval ε, time step γ, individual ID k
Output: Training dataset $Dataset$

```
T_interval = γ, Dataset = Null
While T_interval < ε
    For each point pt_i in Traj_k
        pt_e = Find_nearest(pt_i, T_interval)
        pt_s = pt_i
        For each point pt_j between pt_s and pt_e
            Seq_j = {pt_s, pt_j, pt_e, k}
            Dataset.append(Seq_j)
        End for
    End for
    T_interval += γ
End while
Return Dataset
```

Algorithm 1 is introduced to construct the training dataset with respect to the temporal pattern of the missing data. The proposed algorithm first initializes the training dataset *Dataset* as null and the initial time interval *T_interval* as the referencing time step. Subsequently, we verify if the time interval is less than the maximum time interval ε. If so, we traverse each point in the trajectory in turn as the start point pt_s and search for the point whose recording time is closest to the referencing interval after the start point, which we set as the corresponding end point pt_e. For each pair (start point and end point), we traverse each point pt_j in the trajectory segment as label point and form a triple Seq_j with the start point, the end point, as well as the ID of the individual as one training datum. After all points in the trajectory have been traversed as the start point, we increment the referencing time interval by one temporal unit. The above process is iterated until the time interval *T_interval* is greater than the maximum time interval threshold ε, and the training data *Dataset* are returned as the output of the algorithm. As a consequence of prior operators, there are trajectory segments in the training data *Dataset* with varying time spans and at different times. The parameter settings are calibrated with real-world data in experiments.

4.4. MDP-TR modeling

To implement the trajectory reconstruction method based on the MDP technique, four classic machine learning models were used to reconstruct individual trajectories. As described above, the training dataset are formed as 4-tuples, each of which includes the recorded points pt_{j-1} and pt_{j+1}, the label point pt_j, and the ID of the individual. To estimate the location of the missing point, several selected features are used to describe movement behavior. For trajectory segments, the following features were chosen: the location of pt_{j-1} (x_{j-1}, y_{j-1}), the times of three points $(t_{j-1}, t_j, \text{and } t_{j+1})$, and the location of pt_{j+1} (x_{j+1}, y_{j+1}). The radius of gyration *ROG*, and the trajectory entropy *Ent* were selected to characterize each individual's movement patterns, where *ROG* reflects the range of activity space (typically near the center of the home and work locations for commuters) and *Ent* measures the

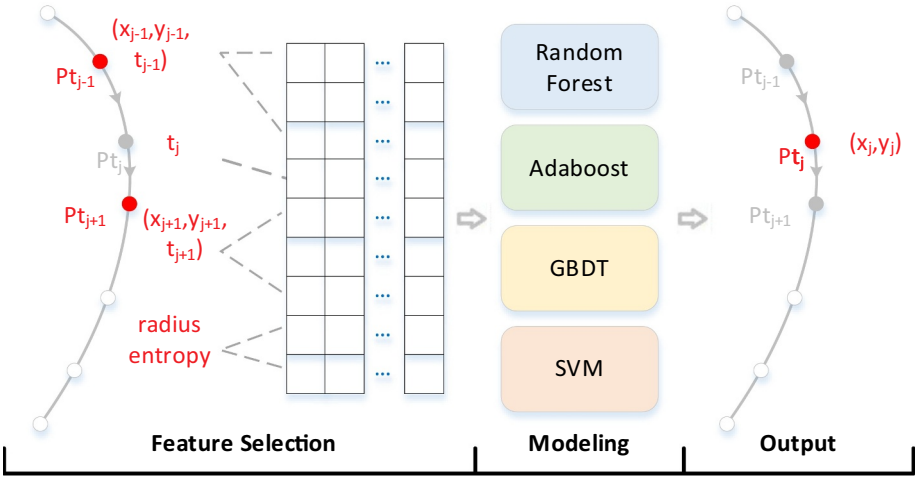

Fig. 5. The architecture of the MDP-TR model.

heterogeneity of movement patterns (Zhao et al., 2016). A larger *ROG* value indicates that the individual has a larger activity area and a larger *Ent* value indicates that the individual has remained in more places and has a more complex movement pattern. These two variables are defined as follows:

$$ROG_k = \sqrt{\frac{1}{m}\sum_{i=1}^{l}(\overrightarrow{pt_i} - \overrightarrow{pt_{\text{cent}}})} \quad (2)$$

$$Ent_k = -\sum_{i=1}^{m} p(i)\log_2 p(i) \quad (3)$$

where m is the number of distinct locations visited by the individual, $\overrightarrow{pt_i}$ represents the i^{th} location of the individual, $\overrightarrow{pt_{\text{cent}}}$ is the mass center of all locations the individual visited ($\overrightarrow{pt_{\text{cent}}} = \frac{1}{m}\sum_{i=1}^{m}\overrightarrow{pt_i}$), and $p(i)$ is the probability that the i^{th} location was visited. That is, the training features can be represented as follows: $[x_{j-1}, y_{j-1}, t_{j-1}, t_j, x_{j+1}, y_{j+1}, t_{j+1}, ROG_k, Ent_k]$. The model is trained with these training features. When performing reconstruction, given the adjacent points pt_{j-1}' and pt_{j+1}', the time of the missing data t_j' which can be any time that we want to know during the time period (t_{j-1}', t_{j+1}'), and the movement characteristics of the individual, the selected features will be formed as a vector that has the same dimensionality and be passed into the trained model. We then obtain the result for the reconstructed missing point. Fig. 5 shows the architecture of the MDP-TR model.

5. Case study

5.1. Data and setup

The performance of the proposed MDP-TR method was investigated using a real-world case study. The data we used was from a mobile communication signaling dataset provided by a major mobile phone operator for scientific research from a city with a population of one million individuals containing 15 million trajectories over a period of 15 workdays. All personally identifiable information was masked. Unlike a traditional CDR dataset, this dataset contains both CDRs and actively generated logs including phone communications, regular updates, periodic updates, cellular handovers, power-ons, and power-offs. Table 1 lists an example of an individual's mobile communication signaling records obtained from the dataset. The coordinate of a cellphone tower is taken as the location of a trajectory point in either the CDRs or signaling records. The dataset was processed to reduce oscillation distortions with a 'ping-pong' suppression (PPS) method proposed by Fiadino, Valerio, Ricciato, and Hummel (2012). For the purpose of verification, we extracted all the CDRs of every individual and stored them in a separate dataset. Therefore, each individual's information appeared in two types of datasets: a CDR dataset and a signaling dataset. An illustration of the difference between the CDR dataset and the signaling dataset is shown in Fig. 6, where the green and red lines indicate the CDR trajectory and the signaling trajectory, respectively, of one individual for one day. In our case study, the trajectory points in the CDR dataset were used as "recorded data" for model training and the trajectory points in the signaling dataset but not in the CDR dataset were used as the "missing data" to evaluate model performance. Fig. 7 shows the probability density function (PDF) and the cumulative distribution function (CDF) of the time spans between two adjacent "recorded data" among individuals.

Table 1
An example of an individual's mobile phone records in the signaling dataset.

Individual ID	Day	Time (t)	Longitude (x)	Latitude (y)	Event type
060F3*****	1	03:38:17	121.58**	31.08**	Periodic update
060F3*****	1	05:54:09	121.58**	31.06**	Periodic update
060F3*****	1	07:38:20	121.58**	31.06**	Call (inbound)
060F3*****	…	…	…	…	
060F3*****	15	23:16:06	121.53**	31.02**	Call (outbound)

The MDP-TR method was coded in the Python programming language. The experiments were performed on a desktop computer with an Intel (R) Core (TM) i7-3770 Processor clocked at 3.40 GHz and having 24 GB main memory, running on the Windows 7 operating system.

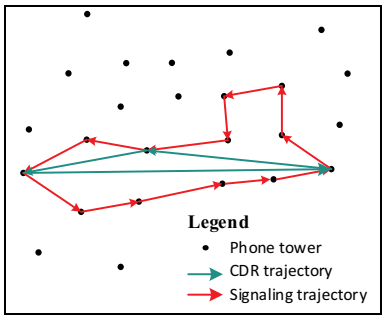

Fig. 6. An illustration of the differences between the CDR and signaling trajectories.

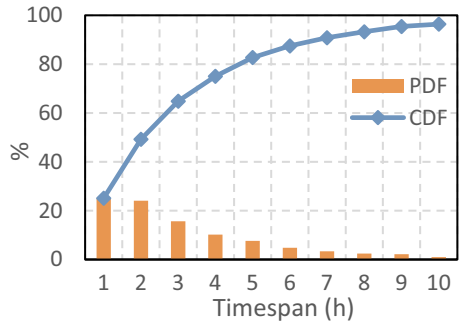

Fig. 7. The cumulative distribution function (CDF) and probability density function (PDF) of time spans among individuals.

5.2. Evaluation metrics

We used the mean absolute error (MAE) to evaluate average performance of the reconstruction models and the standard deviation of error (StDev) metric to evaluate reconstruction model stability. These indicators are defined below.

Given a signaling trajectory, $Traj_k = \langle pt_1, pt_2, \ldots, pt_n \rangle$, and the corresponding CDR trajectory which is reconstructed with the same timestamps, $Traj_k' = \langle pt_1', pt_2', \ldots, pt_n' \rangle$, the MAE is defined as follows:

$$Error_i = |pt_i - pt_i'| \tag{4}$$

$$MAE = \sum_{i=1}^{o} Error_i \Big/ o \tag{5}$$

where $|pt_i - pt_i'|$ is the Euclidean distance between the trajectory location pt_i and the reconstructed location pt_i', o represents the amount of "missing data" which is defined as the trajectory points in the signaling dataset but not in the CDR dataset. Moreover, given the reconstruction error results of the reconstruction error, $Results = \langle Error_1, Error_2, \ldots, Error_o \rangle$, the standard deviation of the error $Stdev$ is defined as follows:

$$Stdev = \sqrt{\sum_{i=1}^{o} (Error_i - MAE)^2 \Big/ o - 1} \tag{6}$$

An example of one individual's reconstruction results was shown in Fig. 8, where the red circles represent the CDR trajectory points, the green line represents the signaling trajectory which was used as ground truth, and the orange line represents the trajectory reconstructed using the method presented in this study.

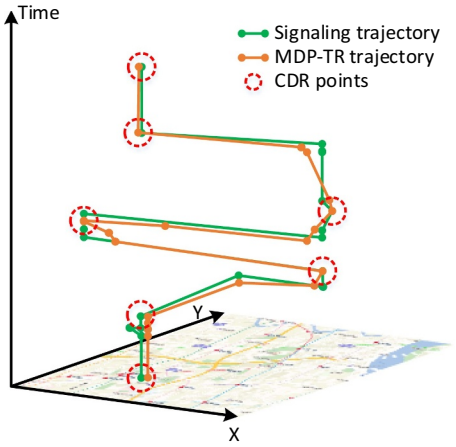

Fig. 8. An example of a reconstruction result.

5.3. Verification of the two algorithms in the MDP technique

As previously mentioned, the MDP technique includes an anchor-point-based clustering algorithm to enhance reconstruction performance by incorporating knowledge of the spatiotemporal similarity between individuals, and an incremental-sliding-window-based data construction algorithm for considering different temporal patterns of missing data. In this subsection, we evaluate the performance of these two algorithms by comparing four strategies: models with the MDP technique (abbreviated as MDP-TR), models with the incremental-sliding-window-based data construction technique only (abbreviated as DC-TR), models with the anchor-point-based clustering technique only (abbreviated as CL-TR), and the original models (abbreviated as Ori-TR). To verify the universality of our method, the trajectories were converted to vectors with the features that were selected in Section 4.4 and then used as input data for the following four machine learning models:

- SVM: Support vector machines construct a set of hyperplanes in a high-dimensionality space. New samples are mapped into the same space and predicted based on the gaps that they fall into (Cortes & Vapnik, 1995).
- RandomForest: The random forest method (RF) constructs a multitude of decision trees and outputs the results by computing the mean of the predictions of each individual tree (Breiman, 2001).
- Adaboost: Adaptive boosting (Ada) is used in conjunction with a set of base learners to improve performance. The outputs of the base learners are combined into a weighted sum to represent the final output (Freund & Schapire, 1997).
- GBDT: The gradient boosting decision tree uses a gradient boosting method to construct a set of decision trees as base learners and outputs the result by computing the sum of the base learners (Friedman, 2001).

Fig. 9 shows the experimental results for MAE and StDev. The results indicate that the MDP technique significantly reduces error and make the reconstruction results more robust with all four models. Furthermore, both the anchor-point-based clustering algorithm and the incremental-sliding-window-based data construction algorithm were helpful in achieving better performance. In addition, out of the four models, we observed that the MDP technique had the best performance with the GBDT model. Therefore, we adopted the GBDT model as the basic model in the MDP-TR method for comparison with the other competing trajectory reconstruction methods.

5.4. Comparison of the MDP-TR method with baselines

To evaluate the advantages of our proposed MDP-TR method over the existing trajectory reconstruction methods, we selected the latest reconstruction method ANN-TR (Z. Liu et al., 2018), and the two most popular methods, linear interpolation and nearest interpolation (Hoteit et al., 2014), as baselines.

- ANN-TR: This method first divides individuals into three groups based on the indicator of the radius of gyration and then feeds movement features of each group into an artificial neural network (ANN).
- Linear interpolation: The results are obtained by joining each pair of consecutive samples of trajectory points with a straight interpolation line. Individuals are assumed to move in a constant direction and at a constant speed during the corresponding time.
- Nearest interpolation: This method consists of using the value of the nearest sampling position in time. Individuals without data are assumed to be stationary.

The performance of the competing methods is shown in Fig. 10. As

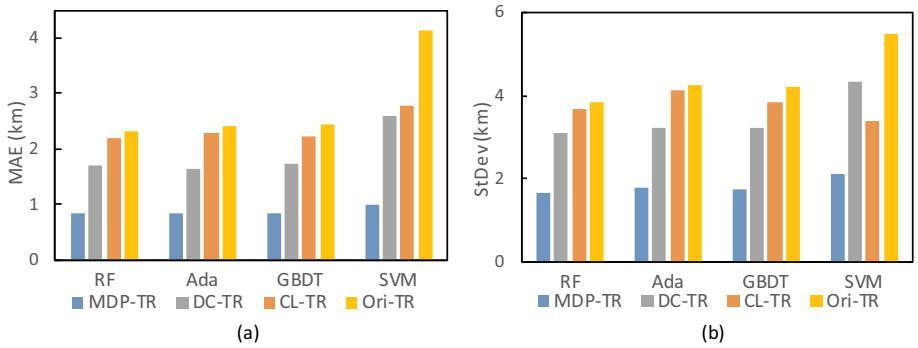

Fig. 9. Verification of the multi-criteria data partitioning (MDP) technique.

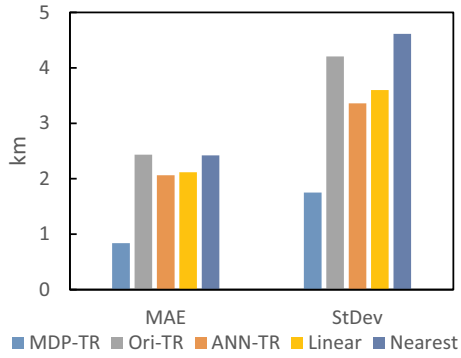

Fig. 10. A comparison with the baselines.

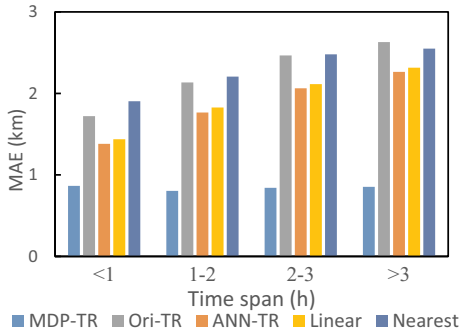

Fig. 11. The effect of the time span.

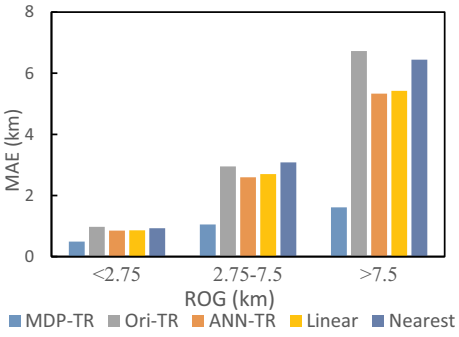

Fig. 12. The effect of the radius of gyration.

shown in the figure, the MAE of MDP-TR was 0.84 km, which was 1.60 km, 1.23 km, 1.27 km, and 1.58 km lower than that for the Ori-TR, ANN-TR, linear interpolation, and nearest interpolation methods, respectively. Moreover, the results of MDP-TR are more robust (i.e., with a minimum StDev) than those of the baseline methods. This improved performance is possibly due to the fact that we accounted for the temporal patterns of the missing data and the similarity among individuals. We also found that when the models were prepared, the reconstruction running times were in milliseconds, which was acceptable for near real-time trajectory reconstruction.

To further demonstrate the advantages of the proposed method, we evaluated the reconstruction accuracy of the proposed MDP-TR method with different time spans and radius of gyration values.

As discussed previously, reconstructions for data with different time spans are quite different. Considering the temporal accuracy of the mobile phone data, we divided the trajectories into four groups with thresholds of 1 h, 2 h, and 3 h time spans, whose results were shown in Fig. 11. We observed that there was a decrease in the average performance of the five methods (i.e., the MAE values increased) with an increase in the time span. This was expected because a greater time span corresponds to a larger choice of alternative movements and increased reconstruction difficulty. However, since the MDP-TR method considers the effects of temporal patterns of missing data during the model training, this method obtained a more robust performance with different time spans. The MAE of the proposed method varied between 0.80 km and 0.87 km with different time spans, which clearly demonstrated the advantages of this method.

The ROG is a concentrated reflection of the subjective willingness and inertial patterns of an individual's mobile activities and reflects the range of activity space. A larger ROG value indicates a wider range of movement. In this study, we categorized individuals into three groups using thresholds of ROG values of 2.75 km and 7.5 km. These thresholds corresponded to the cumulative distribution function values of 60% and 90% in the ROG statistical distribution, respectively. The threshold choice also matched the average travel distance patterns reported in previous studies for intra-urban human mobility, despite a varying peak ROG value due to impacts from city size and shape (Kang et al., 2010, 2012). Fig. 12 showed the results. The five methods obtained good results when the movement range was small. However, with increasing ROG values, the MAE of the proposed method increased from 0.49 km to 1.61 km. Nevertheless, our method still exhibited the lowest reconstruction error for each category of ROG values.

6. Discussion and conclusions

In this study, we presented a novel trajectory reconstruction method using a multi-criteria data partitioning technique, called MDP-TR, to reconstruct the missing records of individuals. For the MDP technique, we adopted an anchor-point-based clustering algorithm to consider the similarity among large-scale low-frequency trajectory data in near real-time and lessen the data hungry problem. We also incorporated an incremental-sliding-window-based data construction algorithm to generate trajectory segments with different time spans and at different times, which allowed the method to learn the movement characteristics

of different temporal patterns of missing data. To evaluate the applicability of the proposed MDP technique, we implemented it using four machine learning models for trajectory location reconstruction. By comparing our method with the competing methods using a large-scale continuous mobile phone dataset of one million individuals over 15 workdays, the advantages of the method in different scenarios were verified. The proposed MDP-TR method outperformed the baseline approaches with respect to the mean absolute error and the standard deviations of error.

By considering the similarity among trajectories and the varieties of temporal patterns of missing data, this method provides an efficient technique to reconstruct the missing locations of individuals. With this more precise reconstruction, future studies could better grasp the status of highly dynamic human movements, which could contribute substantial support for many advanced applications, such as real-time population mapping, short-term traffic forecasting and disaster management. The proposed method could also provide support for sociological studies, such as the analyses of user behavior characteristics and collective behavior patterns. Moreover, when combined with other urban data, such as transportation data and point of interest data, this method could enable the improvement of location-based services (e.g., path optimization, intelligent traffic management, and personalized recommendation) by better understanding of the complex interplay between the urban environment and human activities.

However, there were still several limitations that require further discussion. First, in the MDP-TR method presented here, we only used the Jaccard index and hierarchical clustering to cluster individuals. It is worth noting that using another similarity measure (e.g., the cosine similarity) did not cause a significant change in the clustering results. Therefore, we only used clustering results based on the Jaccard index to further illustrate our analyses and methods in this study. More clustering algorithms and similarity calculation algorithms will be applied to the MDP-TR method to improve its reconstruction performance in future work. Second, considering the frequency of mobile phone used by individuals, the experimental dataset had an average time interval of 30 min. Therefore, fine-grained movement states, such as speed, acceleration, and direction, were not included, which made it nearly impossible to reconstruct missing locations at the minute level. Techniques to extend the proposed MDP-TR method to adapt to a fine-grained trajectory dataset is an interesting but challenging topic for trajectory reconstruction models. In addition, the proposed MDP-TR method only considered the characteristics of the historical trajectories, ignoring the effects of transportation modes, such as traveling by car, by bus, or on foot. Determining an individual's current transportation mode may improve reconstruction accuracy (Shin et al., 2015). Furthermore, in this theoretical methodology-focused study, we ignored the empirical effect of the urban environment. In reality, multiple factors may affect an individual's movements, such as POI distribution, traffic accessibility, and special events (Lindsey, Daniel, Gisches, & Rapoport, 2014; Manley, Orr, & Cheng, 2015). The incorporation of movement behavior is suggested as an extension of the proposed method. It is also worth noting that the data service can facilitate the extension of the concept of calling detail records to cover social networking or web surfing. There might be some possibility for such data becoming available in the future. Our future work should focus on these issues.

Funding

This work was supported by National Natural Science Foundation of China: [Grant Number No. 41771436]; National Key Research and Development Program, China: [Grant Number No. 2016YFB0502104].

Declarations of interest

None.

References

Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies, 18*, 45–54. https://doi.org/10.1016/j.trc.2009.04.011.

Algizawy, E., Ogawa, T., & El-Mahdy, A. (2017). Real-time large-scale map matching using mobile phone data. *ACM Transactions on Knowledge Discovery from Data, 11*, 52. https://doi.org/10.1145/3046945.

Breiman, L. (2001). Random forests. *Machine learning. 45. Machine learning* (pp. 5–32). . https://doi.org/10.1023/A:1010933404324.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Jr., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies, 26*, 301–313.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems, 51*, 70–82. https://doi.org/10.1016/j.compenvurbsys.2015.01.002.

Capela, N. A., Lemaire, E. D., Baddour, N., Rudolf, M., Goljar, N., & Burger, H. (2016). Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants. *Journal of Neuroengineering and Rehabilitation, 13*(1), 5.

Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. *Transportation Research Part B: Methodological, 18*(4–5), 289–299.

Chen, G., Viana, A. C., & Sarraute, C. (2017). Towards an adaptive completion of sparse call detail records for mobility analysis. *2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)* (pp. 302–305). IEEE. https://doi.org/10.1109/PERCOMW.2017.7917577.

Chen, J., & Bierlaire, M. (2015). Probabilistic multimodal map matching with rich smartphone data. *Journal of Intelligent Transportation Systems, 19*, 134–148. https://doi.org/10.1080/15472450.2013.764796.

Chen, L., Özsu, M. T., & Oria, V. (2005). Robust and fast similarity search for moving object trajectories. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (pp. 491–502). ACM.

Chen, Z., Shen, H. T., Zhou, X., Zheng, Y., & Xie, X. (2010). Searching trajectories by locations: An efficiency study. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 255–266). ACM.

Cheng, S. F., & Lu, F. (2017). A two-step method for missing spatio-temporal data reconstruction. *ISPRS International Journal of Geo-Information, 6*, 187. https://doi.org/10.3390/ijgi6070187.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297. https://doi.org/10.1007/BF00994018.

Deng, M., Fan, Z., Liu, Q., & Gong, J. (2016). A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets. *ISPRS International Journal of Geo-Information, 5*, 13. https://doi.org/10.3390/ijgi5020013.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences, 111*, 15888–15893. https://doi.org/10.1073/pnas.1408439111.

Fan, Z., Arai, A., Song, X., Witayangkurn, A., Kanasugi, H., & Shibasaki, R. (2016). A collaborative filtering approach to citywide human mobility completion from sparse call records. In S. Kambhampati (Ed.). *Proceedings of the twenty-fifth international joint conference on artificial intelligence (IJCAI-16)* (pp. 2500–2506). Palo Alto: AAAI Press.

Fiadino, P., Valerio, D., Ricciato, F., & Hummel, K. A. (2012). Steps towards the extraction of vehicular mobility patterns from 3G signaling data. *International workshop on traffic monitoring and analysis* (pp. 66–80). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-28534-9_7.

Ficek, M., & Kencl, L. (2012). Inter-Call Mobility model: A spatio-temporal refinement of call data records using a Gaussian mixture model. *2012 proceedings IEEE INFOCOM* (pp. 469–477). IEEE. https://doi.org/10.1109/INFCOM.2012.6195786.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139. https://doi.org/10.1006/jcss.1997.1504.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Gao, S. (2015). Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition and Computation, 15*, 86–114. https://doi.org/10.1080/13875868.2014.984300.

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS, 17*, 463–481. https://doi.org/10.1111/tgis.12042.

Gao, S., Yan, B., Gong, L., Regalia, B., Ju, Y., & Hu, Y. (2017). Uncovering the digital divide and the physical divide in Senegal using mobile phone data. In D. A. Griffith, Y. Chun, & D. J. Dean (Eds.). *Advances in geocomputation: Geocomputation 2015—The 13th international conference* (pp. 143–151). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-22786-3_14.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., & Trasarti, R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—The International Journal on Very Large Data Bases, 20*(5), 695–719. https://doi.org/10.1007/s00778-011-0244-8.

Gil-Garcia, R. J., Badia-Contelles, J. M., & Pons-Porrata, A. (2006). A general framework for agglomerative hierarchical clustering algorithms. In Y. Y. Tang, S. P. Wang, G. Lorette, D. S. Yeung, & H. Yan (Vol. Eds.), *18th international conference on pattern recognition (ICPR'06). Vol. 2. 18th international conference on pattern recognition (ICPR'06)* (pp. 569–572). Los Alamitos: IEEE. https://doi.org/10.1109/ICPR.

2006.69.
González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779–782. https://doi.org/10.1038/nature06958.
Gurumurthy, K. M., & Kockelman, K. M. (2018). Analyzing the dynamic ride-sharing potential for shared autonomous vehicle fleets using cellphone data from Orlando, Florida. *Computers, Environment and Urban Systems, 71*, 177–185. https://doi.org/10.1016/j.compenvurbsys.2018.05.008.
Hägerstraand, T. (1970). What about people in regional science? *Papers in Regional Science, 24*, 7–24. https://doi.org/10.1111/j.1435-5597.1970.tb01464.x.
Han, X. P., Hao, Q., Wang, B. H., & Zhou, T. (2011). Origin of the scaling law in human mobility: Hierarchy of traffic systems. *Physical Review E, 83*(3), 036117.
Hoteit, S., Secci, S., Sobolevsky, S., Pujolle, G., & Ratti, C. (2013). Estimating real human trajectories through mobile phone data. *2013 IEEE 14th international conference on Mobile data management (MDM 2013). 2. 2013 IEEE 14th international conference on Mobile data management (MDM 2013)* (pp. 148–153). Los Alamitos: IEEE. https://doi.org/10.1109/MDM.2013.85.
Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., & Pujolle, G. (2014). Estimating human trajectories and hotspots through mobile phone data. *Computer Networks, 64*, 296–307. https://doi.org/10.1016/j.comnet.2014.02.011.
Huang, Q., Cao, G., & Wang, C. (2014). From where do tweets originate? A GIS approach for user location inference. *Proceedings of the 7th ACM SIGSPATIAL international workshop on location-based social networks* (pp. 1–8). ACM.
Huang, W., Dong, Z., Zhao, N., Tian, H., Song, G., Chen, G., ... Xie, K. (2010). Anchor points seeking of large urban crowd based on the mobile billing data. In L. Cao, Y. Feng, & J. Zhong (Eds.). *Advanced data mining and applications: 6th international conference (ADMA 2010) proceedings part I* (pp. 346–357). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-642-17316-5_34.
Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*, 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.
Jagadeesh, G. R., & Srikanthan, T. (2015). Probabilistic map matching of sparse and noisy smartphone location data. *2015 IEEE 18th international conference on intelligent transportation systems (ITSC 2015)* (pp. 812–817). Los Alamitos: IEEE. https://doi.org/10.1109/ITSC.2015.137.
Jagadeesh, G. R., & Srikanthan, T. (2017). Online map-matching of noisy and sparse location data with hidden Markov and route choice models. *IEEE Transactions on Intelligent Transportation Systems, 18*, 2423–2434. https://doi.org/10.1109/TITS.2017.2647967.
Kang, C., Gao, S., Lin, X., Xiao, Y., Yuan, Y., Liu, Y., & Ma, X. (2010). Analyzing and geo-visualizing individual human mobility patterns using mobile call records. *Geoinformatics, 2010 18th international conference on* (pp. 1–7). IEEE. https://doi.org/10.1109/GEOINFORMATICS.2010.5567857.
Kang, C., Ma, X., Tong, D., & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1702–1717.
Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS One, 9*(6), e96180.
Lee, J. G., Han, J., & Whang, K. Y. (2007). Trajectory clustering: A partition-and-group framework. *Proceedings of the 2007 ACM SIGMOD international conference on management of data* (pp. 593–604). ACM.
Li, M. X., Lu, F., Zhang, H. C., & Chen, J. (2018). Predicting future locations of moving objects with deep fuzzy-LSTM networks. *Transportmetrica A: Transport Science*. https://doi.org/10.1080/23249935.2018.1552334 in press.
Lindsey, R., Daniel, T., Gisches, E., & Rapoport, A. (2014). Pre-trip information and route-choice decisions with stochastic travel conditions: Theory. *Transportation Research Part B: Methodological, 67*, 187–207. https://doi.org/10.1016/j.trb.2014.05.006.
Liu, K., Gao, S., & Lu, F. (2019). Identifying spatial interaction patterns of vehicle movements on urban road networks by topic modelling. *Computers, Environment and Urban Systems, 74*, 50–61. https://doi.org/10.1016/j.compenvurbsys.2018.12.001.
Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography, 43*, 78–90.
Liu, Z., Ma, T., Du, Y., Pei, T., Yi, J., & Peng, H. (2018). Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. *Transactions in GIS, 22*, 494–513. https://doi.org/10.1111/tgis.12323.
Long, J. A., & Nelson, T. A. (2013). A review of quantitative methods for movement data. *International Journal of Geographical Information Science, 27*(2), 292–318.
Long, Y., & Thill, J. C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems, 53*, 19–35.
Manley, E. J., Orr, S. W., & Cheng, T. (2015). A heuristic model of bounded route choice in urban areas. *Transportation Research Part C: Emerging Technologies, 56*, 195–209. https://doi.org/10.1016/j.trc.2015.03.020.
Mao, Y., Zhong, H., Qi, H., Ping, P., & Li, X. (2017). An adaptive trajectory clustering method based on grid and density in mobile pattern analysis. *Sensors, 17*(9), 2013.
Martínez, L. M., Viegas, J. M., & Silva, E. A. (2009). A traffic analysis zone definition: A new methodology and algorithm. *Transportation, 36*, 581–599. https://doi.org/10.1007/s11116-009-9214-z.
Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems, 5*, 287–301. https://doi.org/10.1080/02693799108927856.
Miller, H. J. (2005). A measurement theory for time geography. *Geographical Analysis, 37*, 17–45. https://doi.org/10.1111/j.1538-4632.2005.00575.x.
Moreno, F., Pineda, A., Fileto, R., & Bogorny, V. (2014). SMOT+: Extending the SMOT algorithm for discovering stops in nested sites. *Computing and Informatics, 33*, 327–342.
Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. *Proceedings of the 2008 ACM symposium on applied computing* (pp. 863–868). ACM.
Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science, 28*(9), 1988–2007.
Ranjan, G., Zang, H., Zhang, Z.-L., & Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review, 16*, 33–44. https://doi.org/10.1145/2412096.2412101.
Shaw, S. L., Yu, H., & Bombom, L. S. (2008). A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS, 12*, 425–441. https://doi.org/10.1111/j.1467-9671.2008.01114.x.
Shin, D., Aliaga, D., Tunçer, B., Arisona, S. M., Kim, S., Zünd, D., & Schmitt, G. (2015). Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems, 53*, 76–86. https://doi.org/10.1016/j.compenvurbsys.2014.07.011.
Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science, 327*, 1018–1021. https://doi.org/10.1126/science.1177170.
Sui, D., & Shaw, S.-L. (2018). Human dynamics in smart and connected communities. *Computers, Environment and Urban Systems, 72*, 1–3. https://doi.org/10.1016/j.compenvurbsys.2018.08.003.
Toohey, K., & Duckham, M. (2015). Trajectory similarity measures. *Sigspatial Special, 7*(1), 43–50.
Wan, L., Gao, S., Wu, C., Jin, Y., Mao, M., & Yang, L. (2018). Big data and urban system model - substitutes or complements? A case study of modelling commuting patterns in Beijing. *Computers, Environment and Urban Systems, 68*, 64–77. https://doi.org/10.1016/j.compenvurbsys.2017.10.004.
Xiao, Z., Wen, H., Markham, A., & Trigoni, N. (2014). Lightweight map matching for indoor localisation using conditional random fields. *IPSN'14: Proceedings of the 13th international symposium on information processing in sensor networks* (pp. 131–142). Piscataway: IEEE. https://doi.org/10.1109/IPSN.2014.6846747.
Xu, Y., Belyi, A., Bojic, I., & Ratti, C. (2018). Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems, 72*, 51–67. https://doi.org/10.1016/j.compenvurbsys.2018.04.001.
Xu, Y., Shaw, S.-L., Fang, Z., & Yin, L. (2016). Estimating potential demand of bicycle trips from mobile phone data—An anchor-point based approach. *ISPRS International Journal of Geo-Information, 5*, 131. https://doi.org/10.3390/ijgi5080131.
Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., & Li, Q. (2015). Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation, 42*, 625–646. https://doi.org/10.1007/s11116-015-9597-y.
Yan, X., Han, X., Zhou, T., & Wang, B. (2011). Exact solution of the gyration radius of an individual's trajectory for a simplified human regular mobility model. *Chinese Physics Letters, 28*(12), 120506.
Yao, X., Zhu, D., Gao, Y., Wu, L., Zhang, P., & Liu, Y. (2018). A stepwise spatio-temporal flow clustering method for discovering mobility trends. *IEEE Access, 6*, 44666–44675.
Yu, H., Russell, A., Mulholland, J., & Huang, Z. (2018). Using cell phone location to assess misclassification errors in air pollution exposure estimation. *Environmental Pollution, 233*, 261–266. https://doi.org/10.1016/j.envpol.2017.10.077.
Yuan, H., Chen, B. Y., Li, Q., Shaw, S.-L., & Lam, W. H. K. (2018). Toward space-time buffering for spatiotemporal proximity analysis of movement data. *International Journal of Geographical Information Science, 32*, 1211–1246. https://doi.org/10.1080/13658816.2018.1432862.
Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2015). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering, 27*, 712–725. https://doi.org/10.1109/TKDE.2014.2345405.
Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior – A case study of Harbin, China. *Computers, Environment and Urban Systems, 36*, 118–130. https://doi.org/10.1016/j.compenvurbsys.2011.07.003.
Yue, Y., Lan, T., Yeh, A. G. O., & Li, Q.-Q. (2014). Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society, 1*, 69–78. https://doi.org/10.1016/j.tbs.2013.12.002.
Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science, 30*, 1738–1762. https://doi.org/10.1080/13658816.2015.1137298.
Zheng, Y. (2015). Trajectory data mining: An overview. *ACM transactions on intelligent systems and technology (TIST). 6. ACM transactions on intelligent systems and technology (TIST)* (pp. 29–). . https://doi.org/10.1145/2743025.