Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 1586010, 12 pages
https://doi.org/10.1155/2021/1586010

Research Article

Refining Sparse Cell-ID Trajectory of Public Service Vehicles by Spatiotemporal Modelling

Kemin Zhu,[1] **Junli Liu**,[1] **Xianfeng Song**,[1,2,3] **Weifeng Wang**,[1] **and Hao Chen**[4]

[1]*College of Resources and Environment, Chinese Academy of Sciences, Beijing 100049, China*
[2]*Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100049, China*
[3]*Key Laboratory of Quantitative Remote Sensing Information Technology, Chinese Academy of Sciences, Beijing 100049, China*
[4]*School of Urban and Environmental Sciences, Huaiyin Normal University, Huai'an 223001, Jiangsu, China*

Correspondence should be addressed to Xianfeng Song; xfsong@ucas.ac.cn

Received 14 October 2019; Revised 20 October 2020; Accepted 15 January 2021; Published 27 January 2021

Academic Editor: Rocío de Oña

Mobile phone data have become a critical data source for transportation research. While a cell-id trajectory was routinely reorganized by International Mobile Subscriber Identity (IMSI), it potentially allows to analyze transportation behaviors and social interaction of total population, with a full temporal coverage at low cost. However, cell-id trajectory is often sparse due to low reporting frequency and uncertainness of mobile holders' position. So, the cell-id trajectory refinement has been recognized as challenging work to further facilitate trajectory data mining. This paper presents a comprehensive approach to identify cell-id trajectories of public service vehicles (PSVs) from large volume of trajectories and further refines these cell-id trajectories by a heuristic global optimization approach. The modified longest common subsequence (LCSS) method is used to match a cell-id trajectory and a public transportation route (PTR) and correspondingly calculates their similarities for determining whether the trajectory is PSV mode or not. Taking full advantages of the nature of a PSV tends to move on the PTR in uniform motion to meet a prescript visit to stops, a heuristic global optimization approach is deployed to build a spatiotemporal model of a PSV motion, which estimates new locations of cell-id trajectories on the PTR. The approach was finally tested using Beijing cellular network signaling datasets. The precision of PSV trajectory detection is 90%, and the recall is 88%. Evaluated by our GNSS-logged trajectories, the mean absolute error (MAE) of refined PSV trajectories is 144.5 m and the standard deviation (St. Dev) is 81.8 m. It shows a significant improvement in comparison of traditional interpolation methods.

1. Introduction

Cellular network-based data are emerging as a great data source for urban transportation application due to the advantage in the large geographic coverage of cellular networks and the comprehensive penetration in a population [1–3]. Generally, cellular network-based data collected by mobile network operators can be reorganized into a mobile phone user's trajectory formed as a sequence of time-stamped cell-ids (i.e., cell-id trajectory), illustrating motion characteristics of corresponding mobile objects [4].

The location in cell-id trajectory at a particular time is assigned with the coordinate of an occupied base transceiver station (BTS). The spatial resolution of cell-id trajectory data depends on the service radius of each BTS, which varies in different areas, e.g., of hundred meters in metropolitan cities, and several kilometers in rural areas [5]. Meanwhile, the records of a cell-id trajectory are collected in a relatively long-time interval, which inevitably results in the problem of trajectory discontinuity or sparsity [6–10]. Therefore, due to the inconformity between the spatiotemporal sparsity of the cell-id trajectory data and the requirement on fine-scale footprints, the refinement of cell-id trajectories becomes a research hotspot due to its extensive application prospects.

The trajectory refinement that refers to filling the spatiotemporal gaps in the data is an approach to mitigate the sparsity of time-series trajectories [11]. Existing refinement method based on mobile phone location data mainly

includes an interpolation-based method and a map-matching-based method. The former mainly uses spatial-temporal correlations among records to interpolate missing points. During interpolation, trajectory points are calculated based on the distances and time spans between each missing point and its contextual points by an estimation function (e.g., nearest-neighbor function, linear function, or Gaussian function). Ficek and Kencl [12] proposed an intercall mobility model which combines Gaussian mixtures to refine CDRs. Hoteit et al. [13, 14] compared the reconstruction performance of various interpolation methods (linear, cubic, nearest, and spline interpolations) on trajectories with different sampling interval and radius of gyration. Yu et al. [15] simply used a spatially-linear-interpolated method to estimate exposure in air pollution of cell phone user. Csaji et al. [16] interpolated between home and office to infer user's location and analyze the population distribution. These refinement methods might meet the requirement for pollution exposure or census estimation; however, it is unfeasible for transportation study. Perera et al. [17] provided a method to compute the location of phone user within a cell, but it requires extra speed information which is generally unavailable in cell-id data.

As for map-matching-based methods, the basic assumption is that vehicle movement behaviors always occur along road networks. Thus, a sequence of trajectory points can be aligned to a sequence of road segments to form a complete path. Hidden Markov model (HMM) is the most popular one among map-matching-based methods. Jagadeesh and Srikanthan [18] used a pretraining route choice model to generate partial map-matched paths and identify the most likely one. Another HMM model proposed by Jagadeesh and Srikanthan is in [19] which considered the tailored transition probabilities for the type of data. Algizawy et al. [20] used HMM to generate a road-level traffic density, at an hourly granularity, for each mobile trajectory. Xiao et al. [21] used contextual relationships between trajectory points as features of the CDR trajectories in a conditional random field model to reconstruct individual trajectories. Chen et al. [22] proposed two approaches for completing CDRs adaptively to reduce the sparsity and mitigate the problems the latter raises. However, the basic assumption that underlies the map-matching method is questionable. That is, individuals in urban space can travel by public transportation, which limits the performance of such methods.

This paper aims to refine sparse cell-id trajectory of public service vehicles (PSVs) by combining vehicle transport model and mobile cell-id trajectories, in which an LCSS-based SVM classifier took full advantage of the similarity between cell-id trajectories and designed public transportation routes (PTRs) to separate PSVs from those with other transport modes (e.g., walk or private car) in a large-scale unlabeled trajectory dataset. Then, each trajectory of PSV was modeled to fit with its mobile behaviors as much as possible at stops, junctions, and roads and be consistent to a spatial cell cover and bus travel speed by a heuristic global optimization. We evaluated our proposed trajectory refinement method by using an encrypted cell-id trajectory dataset and a GNSS-logged bus trajectory collection. The results show that our approach delivers a state-of-the-art achievement in refining cell-id trajectories.

2. Method

2.1. Conceptual Model. The cell-id trajectory explored in this paper is a sequence of time-stamped cell-ids. Each cell-id is a unique identification of a base transceiver station (BTS) with geographical coordinates and sectoral signal transmission coverage. So, we can imagine that a refined cell-id trajectory must occur within the overlap area among a road network and a BTS sectorial area. Figure 1 gives an overview of the architecture of our approach, including two phases: PSV trajectory detection and trajectory refinement or reconstruction.

PSV cell-id trajectory detection is to identify PSV-generated trajectories from huge volume of cell-id trajectories with various transportation modes, i.e., walk, cycle, and private car. An LCSS alignment algorithm is used to match a cell-id trajectory to a public transportation route (PTR) according to a number of PTR-nearby BTS sectoral coverages and chronological order of timestamps. LCCS generates a sequence of corresponding points (also called anchor points), in which an anchor point is represented to two different locations on the PTR route and cell-id trajectory, respectively.

The similarities calculated based on the LCSS sequence are sufficient conditions to recognize PSV trajectory by a support vector machine classifier. Anchor points are initial inputs to heuristic optimization model for further estimating precise locations of anchors on PTR routes, and consequently, high-quality trajectories are generated by interpolating among optimized anchor points.

2.2. PSV Cell-ID Trajectory Detection. The PSV such as buses, subways, and trams always run following the transportation mode with fixed routes and prescript schedules, providing transportation services for public passengers. To identify cell-id trajectories from mass datasets, it is essential to establish a set of specific measures to quantify the correlation between cell-id trajectories and the PTR. Based on LCSS, the longest common sequence among a cell-id trajectory and PTR is matched and thus a set of similarity measures are proposed to measure the spatiotemporal correlation between them. Taking use of the similarity measures, a SVM binary classifier is deployed to recognize cell-id trajectories with the PSV mode.

2.2.1. BTS Sector. In a cell-id trajectory, a phone holder's location is roughly represented as the installation position of the cell-id marked BTS. It is not the identical location of the phone holder as the BTS actually covers a large geographical area. It is impossible to catch the exact location of mobile users as they can be anywhere inside the mobile network cell.

There are two popular mobile network cell models. Most researches represent mobile network cells as Voronoi areas centered on BTSs (see Figure 2(a)). In this work, BTS sectors

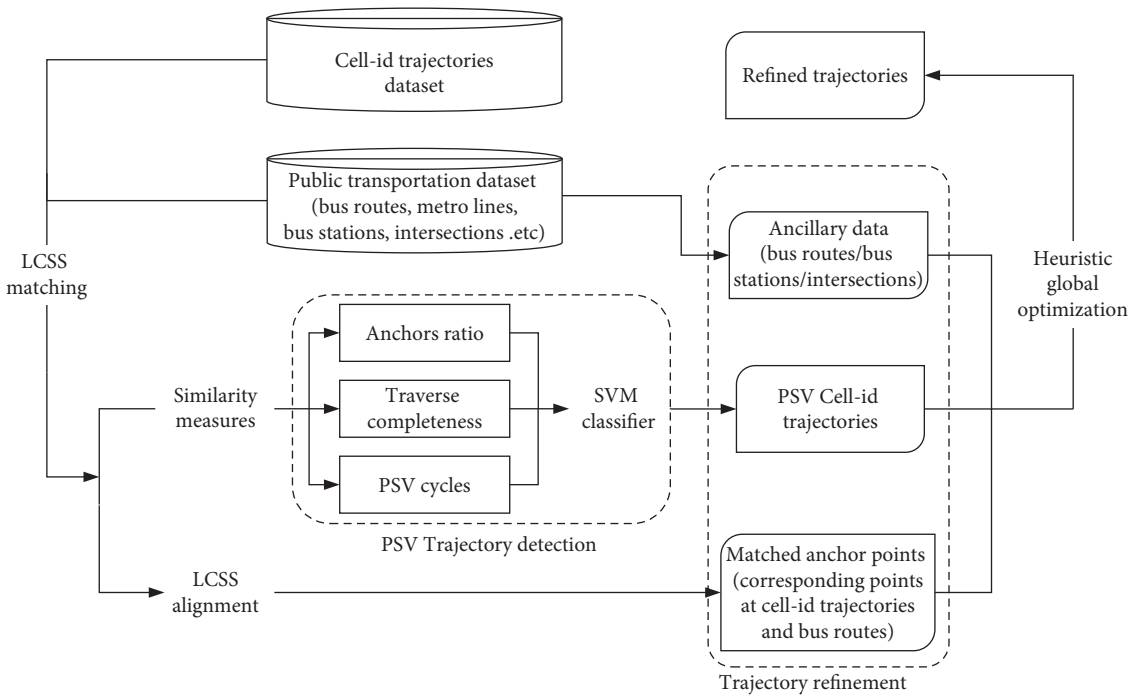

Figure 1: Diagram of sparse cell-id trajectory refinement.

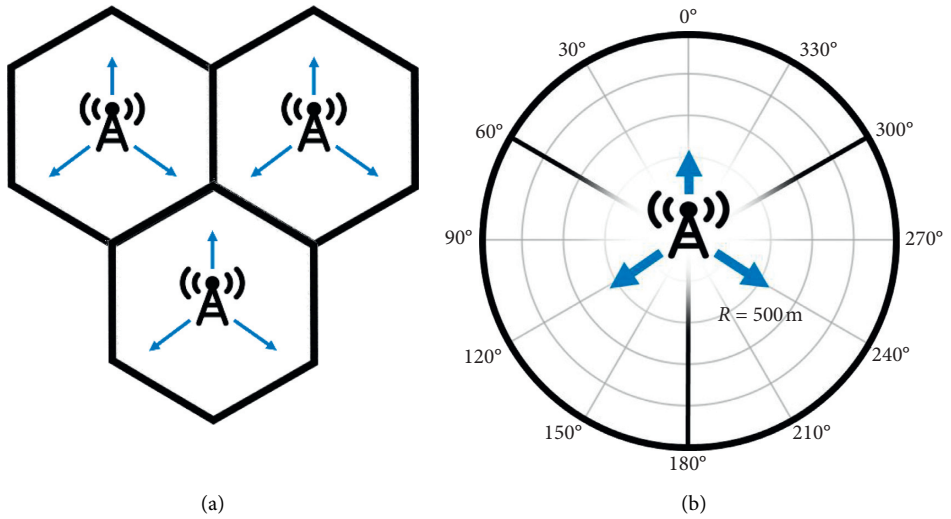

Figure 2: (a) Voronoi centered on BTS. (b) Trisector BTS with 3 antennas.

are used that offer a finer spatial scale due to limiting a mobile phone location to only parts of the cell, as shown in Figure 2(b). Generally, there are three antenna sectors per BTS and the sector's orientation is same as the direction of BTS antenna. We construct sectors with an average diameter of 500 meters in urban areas.

2.2.2. LCSS Matching. Due to the uncertainness of spatial positioning and irregular time interval logging at a cell-id tracking point, it is difficult to match a PSV cell-id trajectory to a PTR route. An LCSS alignment algorithm that originates in the field of string matching, where two strings are given to find characters that appear left-to-right, not necessarily consecutively, in both strings [23], is applied to find the longest common subsequence of two sequences. The longer an LCCS is, the higher the probabilities that the trajectory was generated by a PSV.

The LCCS is extended to support periodic matching in this work. Following pilot studies [24, 25] that processed cell-id trajectories as strings, we discretized a continuous PTR route into a cyclic sequence of discrete points with an

ALGORITHM LCSS matching with Dynamic programming

INPUT: A spatial-temporal trajectory T and a route sequence R

OUTPUT: The optimal anchor points sequence $LCSS$

1: $row = \|T\| + 1, col = \|R\| + 1, row \times col\ matrix\ D, \varepsilon = \|R\|/2, loc = 0, cyc = 0$
2: **for** each entry $D[i,j]$ in D **do**
3: $D[i,j] = 0$
4: **for** i in 1 to row **do**
5: **if** $j + ext > col$ **then**
6: $R.append(R_{(j,j+\varepsilon)})$
7: $col += \varepsilon$
8: **for** j in 1 to col **do**
9: $S = D[i-1, j-1] + Sim(T.p_i, R.q_j)$
10: $D[i,j] = \max(D[i-1,j], D[i,j-1], S)$
11: $LCSS = \emptyset, i = \|T\|, col = \|R\|$
12: **while** $i > 0$ and $j > 0$ **do**
13: **if** $D[i,j] = D[i-1,j-1] + Sim(T.p_i, R.q_{j\%\|R\|})$ **then**
14: Insert $(T.p_i, R.q_{j\%\|R\|})$ into $LCSS$, $i--$, $j--$
15: **else if** $D[i,j] = D[i-1,j]$ **then**
16: Insert $(T.Pnt_{i-1}, "-")$ into $LCSS$, $i--$
17: **else if** $D[i,j] = D[i,j-1]$ **then**
18: Insert $("-", R.q_{j\%\|R\|})$ into $LCSS$, $j--$
19: **while** $j > 0$ do
20: Insert $("-", R.q_{j\%\|R\|})$ into $LCSS$, $j--$
21: **while** $i > 0$ do
22: Insert $(T.p_i, "-")$ into $LCSS$, $i--$

Figure 3: LCSS extension for periodic matching.

interval of 50 meters. Given a cell-id trajectory T and a PSV route R,

$$T = \{(t_i, p_i)\}, \quad i = 1, \ldots, n,$$
$$R = \{q_j\}, \quad j = 1, \ldots, m, \tag{1}$$

where p_i is the ith point on trajectory T, t_i is the timestamp of p_i, and q_j is the jth point on route R.

Two points p_i and q_j may be considered to be matched if q_j located within the BTS sector of p_i, and it can be represented as

$$m(p_i, q_j) = \begin{cases} 1, & q_j \in S_{p_i}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where S_p is the BTS sector of p_i.

Let $T_{(n)}$ denote the first n points of trajectory T and $R_{(m)}$ denote the first m points of R, the LCSS between T and R.

$$\text{LCSS}(T_{(n)}, R_{(m)}) = \begin{cases} 0, & n = 0, \\ \max\begin{pmatrix} \text{LCSS}(T_{(n-1)}, R_{(m)}) \\ \text{LCSS}(T_{(n)}, R_{(m-1)}) \end{pmatrix} + m(p_m, q_n), & n \neq 0. \end{cases} \tag{3}$$

Dynamic programming algorithm deployed to discover optimal alignment for T and R is shown in Figure 3. A matrix D is created to save the LCSS for every subsequence pair of T and R. $\|T\|$ and $\|R\|$ are the point numbers of T and R, respectively. Considering a PSV periodically moves along PTR, R is designed as a periodic cyclic sequence. ext is the search space on R for T. p_i to find matched $R.q_j$. The entries of D are gradually filled as the dynamic programming proceeds (lines 4–10), and the last entry stores the LCSS of aligning T and R. Finally, we decode D to find all the aligning cell-id log pairs of the optimal alignment (lines 12–22). The extended LCSS model can be viewed as a modified version of the models [26, 27], which not only finds the longest common subsequence in terms of the accumulate number of matched anchors LCSS(T, R) but also the cycle numbers of the cell-id trajectory.

2.2.3. Similarity Measures

(a) Anchors Ratio. An often used similarity measure for an LCSS matching is calculated as the ratio between the number of points in LCSS and the number of cell-ids in original cell-id sequences size [26, 27], called anchors ratio in this work.

$$\mathrm{SC}(T,R)=\frac{\|\mathrm{LCSS}(T,R)\|}{\|T'\|}, \tag{4}$$

where $\mathrm{LCSS}(T,R)$ is the number of the matched anchor point pairs, and T' is a subset of T that is the common part between T and R. Using T' instead of T is because T contains points beyond on-duty period..

(b) Traverse Completeness. Traverse completeness refers to the ration between the length of the LCSS and the length of a PTR. Anchor points of $\mathrm{LCSS}(T,R)$ split T and R, resulting in a set of subtrajectories $\{\mathrm{ST}_1,\ldots,\mathrm{ST}_i,\ldots,\mathrm{ST}_n\}$ and a set of subroute $\{\mathrm{SR}_1,\ldots,\mathrm{SR}_i,\ldots,\mathrm{SR}_n\}$. Applying Hausdorff distance [28] to determine whether $(\mathrm{ST}_i,\mathrm{SR}_i)$ is matched with each other, all geometrically closed pair $(\mathrm{ST}_j,\mathrm{SR}_j)$, $j=1,\ldots,m$ could be regarded as traversed partial on the PTR; correspondingly, traverse completeness TC is defined as follows:

$$\begin{aligned}\mathrm{TC}(T,R)&=\frac{L(\overline{\mathrm{MSR}})}{L(R)},\\ \overline{\mathrm{MSR}}&=\left(\mathrm{SR}_1\bigcup\ldots\bigcup\ \mathrm{SR}_j\bigcup\ldots\mathrm{SR}_m\right),\end{aligned} \tag{5}$$

where $\overline{\mathrm{MSR}}$ is the union of matched subroutes SR_j, $j=1,\ldots,m$, $L(\overline{\mathrm{MSR}})$ is the length of union-routes $\overline{\mathrm{MSR}}$, and $L(R)$ is the total length of route R.

(c) PSV Cycles or Round-Trips. PSV cycles denote to the cycles of periodic matching of cell-id trajectory on the PTR, indicating how many round-trips the PSV run along the route.

$$\mathrm{CYC}\ (T,R)=\frac{m_e-m_s}{L(R)}, \tag{6}$$

where m_s and m_e are the mileages of the first and last anchor points, respectively, m_e-m_s represents the total distance during on-duty time, and $L(R)$ is the total length of route R.

2.3. Trajectory Refinement

2.3.1. Trips Partition. A PSV cell-id trajectory often includes several round-trips along a PTR route, each alternating with a long-time stay at terminal stations as drivers have access to toilet facilities at rest, fuel, and food establishments. The stop time cannot be directly obtained from original data, so we introduced a spatiotemporal kernel density estimation (STKDE) method to identify these stays and therefore enable to partition trips. Trajectory refinement or reconstruction is thus able to be performed trip by trip because those short trajectories generated at terminal stay time are eliminated.

We first transform the cell-id trajectory into a time-distance relationship by calculating the distance between start station and each BTS. Then, the kernel density along the distance axis is estimated as

$$\widehat{f}(d)=\frac{1}{h_d}\sum_i K_s\left(\frac{d-d_i}{h_d}\right), \tag{7}$$

where m is the accumulated mileage of the PSV, K_s is a kernel function for the spatial domain, and h_s is the spatial bandwidth. Trajectory point is weighted on a univariate kernel density function K_s as follows:

$$K_s(u)=\begin{cases}\dfrac{2}{\pi}\left(1-u^2\right), & u^2<1,\\ 0, & \text{otherwise.}\end{cases} \tag{8}$$

The kernel density $\widehat{f}(d)$ is estimated based on the travel duration when the PSV pass through the spatial domain. Points on which the PSV has met traffic congestion, junctions, or stops often have high-density values. In particular, the long-time stay at terminal stations causes an extremely high-density peak, which can be used to partition trips [29].

2.3.2. Heuristic Global Optimization. For a round-trip matching, we deploy a heuristic global optimization approach to estimate the precise locations of anchor points on the route. It tends to match a time-stamped point sequence to a monotonically increased mileage sequence. For each anchor point of the LCSS from a cell-id trajectory and a PTR, it must satisfy (a) locating within the intersection of the PTR and BTS sectors, (b) having a longer mileage than previous points' one, and (c) having new location nearby the initial.

A heuristic optimization model, as a commonly used model on finding approximate global optima problems, is used to search new location of anchor points naturally. Assuming a PSV always tends to move on a PTR in uniform motion to meet a prescript visit to stops, an objective function is defined as equation (8) to minimize the standard deviation of bus speed along a route in which the speeds among two consecutive anchors totally depend on their locations and time interval.

$$\begin{aligned} &\text{minimize:} \quad F(m_i) = \sum_{i}^{n-1} \left[\frac{m_{i+1} - m_i}{t_{i+1} - t_i} - \frac{1}{n-1} \sum_{i}^{n-1} \left(\frac{m_{i+1} - m_i}{t_{i+1} - t_i} \right) \right]^2, \\ &\text{subject to:} \quad m_i < m_{i+1}, \\ &\qquad\qquad\quad lb_i < m < ub_i, \end{aligned} \tag{9}$$

where m_i is the mileage of the ith anchor along the PTR, t_i is the timestamp, n is the total number of anchor points, and lb_i, ub_i are the upper and lower limits of the decision variables m_i being optimized.

As illustrated in Figure 4, by iteratively adjusting the location of anchor points in a search space, a minimum objective function value is reached by leveraging dwelling times on bus stops and cross-roads besides the above-mentioned constraints. The steps of the method are listed as follows.

3. Study Area and Datasets

3.1. Study Area. The study was conducted at Huilongguan town to evaluate the proposed public transit mode detection and cell-id trajectory refining method. Huilongguan town is one of the largest townships in the northern part of Beijing, China. The town has a total population of about 450,000 people and an area of 34.5 square kilometers. Being one of the most populated residential areas in Beijing, choked traffic has been an archenemy to urban transportation system in this area. As shown in Figure 5, the town streets and road maps were sourced from OpenStreetMap and total 20 bus routes and related more than 400 bus stops covering this area were also retrieved from Beijing Public Transport Corporation (BPTC).

3.2. Datasets. Cellular phone network signaling datasets, covering the town extension area on early August, 2016, were collected, about 670,000 mobile phone trajectories, 240 million records with an average phone-station interaction interval of 280 seconds. Each record includes timestamp (TS), International Mobile Equipment Identity (IMEI), tracking area code (TAC), and cell identity (CI), representing an interaction event between a mobile phone (IMEI) and a base station (TAC plus CI) at a dedicated time (TS). All private information in mobile phone datasets has been encrypted to protect privacy.

Ground truth datasets were collected by deploying an android device-based cell signal monitor program. Following GNSS positioning (longitude, latitude, and time), cellular towers information along bus routes was also acquired, including cellular network (GPRS/EDGE/UMTS/LTE), current cell identity (CID), current area identity (LAC/RNC/TAC), signal strength (RSSI and RSRP for LTE networks), and cells that were used by the mobile device. This tracking dataset was logged with a sampling interval of 1 second and a GPS positioning error of 5–10 meters. It is mainly used to calibrate and validate our proposed model.

3.3. Data Preprocessing. Ground truth datasets were reorganized as follows. First, the information about bus round-trips was extracted from GNSS-measured bus trajectories. Then, bus cell-id trajectories were prepared by resampling raw cellular tacking datasets with a long-time interval of 280 seconds to be consistent with our big cellular collection in 2016.

As shown in Figure 6, a subset of ground truth datasets at route no. 307, totally 3,8273 GNSS points and 136 cell towers of BTS, were illustrated, among which each BTS might be deployed many times when our android devices passed through its covered area. Figure 6 displays both a GNSS trajectory and a cell-id trajectory from 8 : 00 to 20 : 00 within 4 separate round-trips. GNSS points overlap bus route very well but cell-id points are quite poor.

4. Results

4.1. PSV Cell-ID Trajectory Detection by LCSS. Applying our revised LCSS algorithm to register cell-id trajectories datasets to 20 bus routes in Huilongguan, the similarities between trajectories and bus routes were calculated. Total 843 candidate cell-id trajectories with 16,732 time-stamped track points (or towers) were identified with an anchor ratio threshold of 0.2. In this work, the cell-id trajectories that are from bus drivers or conductors are called "PSV mode." By human interpretation, 456 cell-id trajectories (phone holders) were labelled as "PSV mode" and the rest were "non-PSV mode." The statistics of similarity measures of PSV mode trajectories is listed in Table 1.

Support vector machine (SVM), which may maximize the margin by separating two classes of samples, was used to identify PSV cell-id trajectories from others in this work. K-fold cross-validation was used for SVM parameters tuning such that the model with most optimal value of hyper-parameters can be trained. The interpreted candidates are divided into 5 folds, out of which four folds are for training and one for testing. The results of five repeated classifications are shown in Table 2. For PSV cell-id trajectory detection, the precision is about 90% while the recall is from 88.60% to 92.32%. For non-PSV mode detection, the precision is around 88%, and the recall is from 88.37% to 93.54%.

4.2. Trajectory Refinement by Heuristic Optimization. Once the trajectories with a PSV mode were identified, our proposed heuristic optimization method is used to determine the precise location of anchor points on a bus route at a specific timestamp by concerning its corresponding BTS sector and other contextual information. To evaluate the

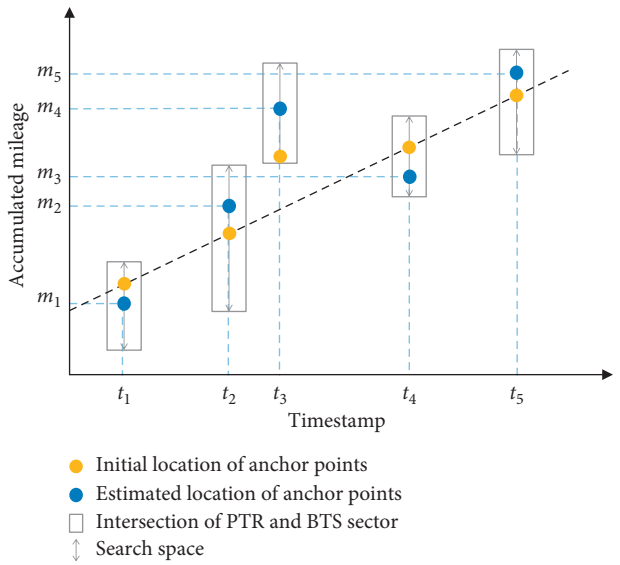

Figure 4: Heuristic optimization of the localization of anchor points.

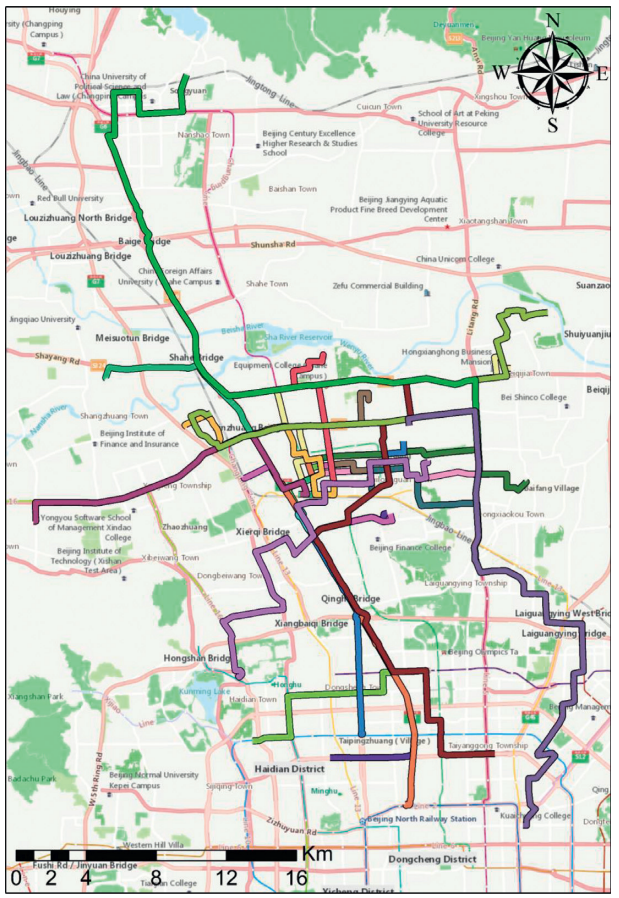

Figure 5: Huilongguan town and bus routes.

performance of our heuristic globe optimization approaches, the ground truth datasets were deployed. As the result of trajectory matching shown in Figure 7, 81 out of 136 BTS

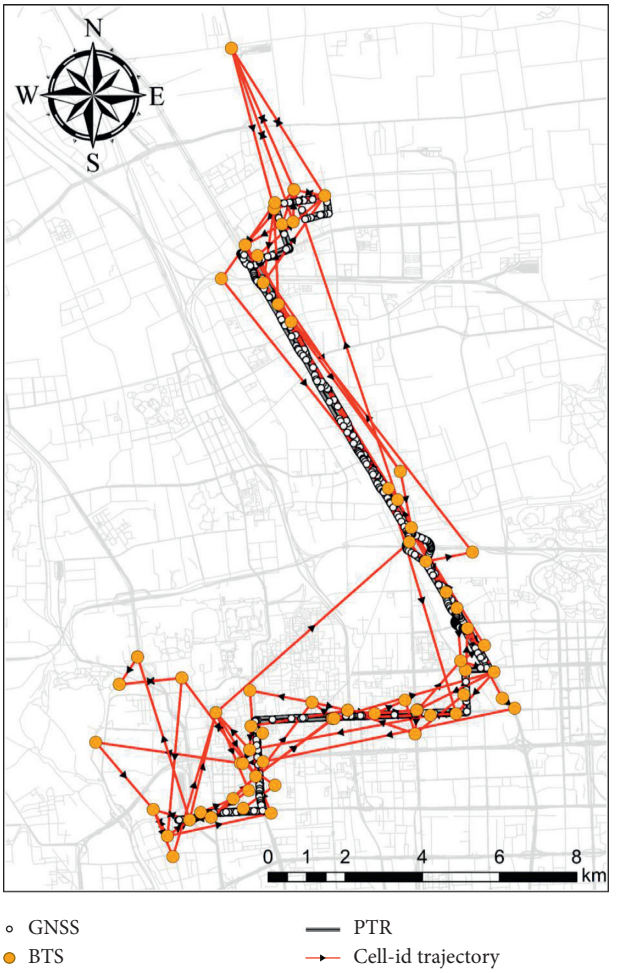

Figure 6: Ground truth data (GNSS vs cell-id).

along bus route no. 307 were detected that they have counterparts anchor points existing on the route.

Five optimization algorithms including particle swarm optimization (PSO), augmented Lagrangian (AL), compass search (CS), and artificial bee colony (ABC) were tested to solve the optimization problem of spatiotemporal modelling of a cell-id trajectory. For our PSV's location optimization, it is actually a continuous, constrained, and single-objective problem. That is, an anchor point of a BTS needs to be found in optimization space which is the common segment part of the BTS sector and the bus route.

A python library PyGMO was deployed for implementing this iterative process, and all results were evaluated by the indicators of MAE (mean absolute errors) and St. Dev (standard deviation) by referring to ground truth GNSS tracking points. As shown in Figure 8, the compass search algorithm achieved the best performance in comparison of others. The estimated locations of anchor points at their BTS-communication time have the smallest errors. Based on the optimized cell-id trajectories, we further make a spatiotemporal interpolation among anchor points using method in [13] and reconstruct a high-quality PSV trajectory.

Table 1: Similarity measures of PSV cell-id trajectories.

	Mean	St. Dev
Anchor ratio	0.79	0.13
Traverse completeness	0.89	0.12
PSV cycles	3.24	0.83

Table 2: Classification results.

		PSV	Non-PSV	Total # of true T	Recall (%)
Test 1	PSV	421	35	456	92.32
	Non-PSV	41	346	387	89.41
	Total # of est. T	462	381	843	90.98
	Precision	91.13%	90.81%	–	–
Test 2	PSV	415	41	456	91.01
	Non-PSV	25	362	387	93.54
	Total # of est. T	440	403	843	92.17
	Precision	94.32%	89.83%	–	–
Test 3	PSV	404	52	456	88.60
	Non-PSV	39	348	387	89.92
	Total # of est. T	443	400	843	89.21
	Precision	91.20%	87.00%	–	–
Test 4	PSV	402	54	456	88.16
	Non-PSV	41	346	387	89.41
	Total # of est. T	443	400	843	88.73
	Precision	90.74%	86.50%	–	–
Test 5	PSV	412	44	456	90.35
	Non-PSV	45	342	387	88.37
	Total # of est. T	457	386	843	89.44
	Precision	90.15%	88.60%	–	–

Figure 9 presents the iterative process of a trip of PSV cell-id trajectory data using compass search algorithm. A monotonic basin hopping (MBH) meta-algorithm is applied.

To avoid local optimization and accelerate convergence velocity of iteration, a monotonic basin hopping meta-algorithm is applied in the abovementioned optimization processes. The optimization space threshold λ roughly set to 500 m to avoid an explicitly unreasonable location was tested in the iteration. As shown in Figure 10, the objective decreases quickly at the beginning and finally converges at minimization after 120,000 iterations. Generally, an acceptable result could be reached after 8,000 iterations.

4.3. Comparison with Other Spatiotemporal Interpolations. To evaluate the advantages of our proposed method, we selected three most popular trajectory reconstruction methods, nearest sampling (nearest), linear interpolation (L), and cubic hermit interpolation (CH), [13] to make a comparison. These popular interpolations have two models, constrained (C) and unconstrained (U), depending on whether ancillary transportation line datasets are used or not. Calculating the difference between the estimated location and the GNSS measured location of a tracking point at a specific timestamp, the performances of the abovementioned methods are shown in Figure 10, among which our method (compass search algorithm) has the smallest error on its refined cell-id trajectory.

The box of CS errors is compact with a small median of 166 m. Among traditional interpolation methods, constrained linear interpolation seems introducing a relatively good result with a median of 207 m, but the range of errors widely varies. Moreover, CS is also robust and the better performance is due to the fact that we accounted for the unique transportation model of PSV and the constraints of public transport line.

5. Discussion

5.1. Shift in LCSS Matching. Assuming a bus route crosses a BTS sector that is a wireless-signal coverage area (Figure 11), it leads to a piece of road intersection on which a mobile holder (a bus driver) moves when the phone communicates with the corresponding BTS at a particular time. The end-points of the interaction may be called, entry point and exit point. All communication events with this BTS must happen on the interactions with big probabilities.

The revised LCSS for trajectory matching is a forward matching process. This process is able to find the longest common subsequence, but cannot guarantee an appropriate anchor point that is expected to be the identical location of the bus when moving on roads at a specific time. Among the discretized points on the intersection part, the closer a point is to the entry, the higher the probabilities of the point are to be matched. This is because the effect of mismatch and process variation results in shifting in the process of LCSS.

Our heuristic optimization with compass search algorithm further adjusts the initial positions of cell-id trajectory points within the intersection part and estimates appropriate positions to match a PSV-mode transportation along the bus route. In comparison of four GNSS trajectories, the average distance between initial points and estimated points is about 197 m with a deviation of 57 m.

5.2. How a Time Interval Affects the Spatiotemporal Modeling in Trajectory Refinement? From the comparison of heuristic optimization and traditional interpolations in the above section, our proposed method did improve the accuracy of trajectory refinement. Moreover, the inherent quality of cell-id trajectories, particularly the time intervals of tracking points on a cell-id trajectory, also has big effects on the refinement.

Based on the ground truth datasets collected in four round-trips at bus route 307, the cellular signaling data originally logged with one second interval were resampled into a number of trajectories with intervals ranging from 10 s to 600 s. The BTS (cell-id) number in these resampled

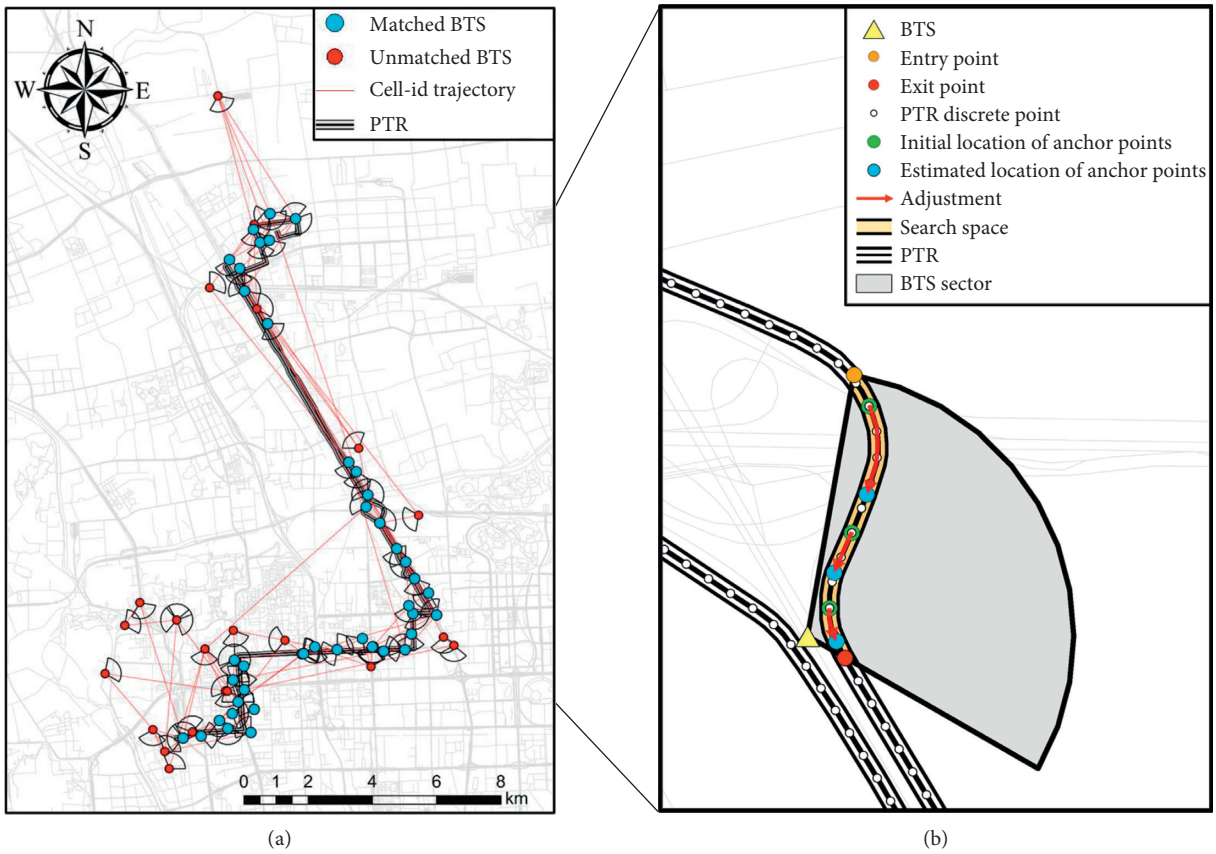

Figure 7: Anchor point's search space in a matched BTS sector.

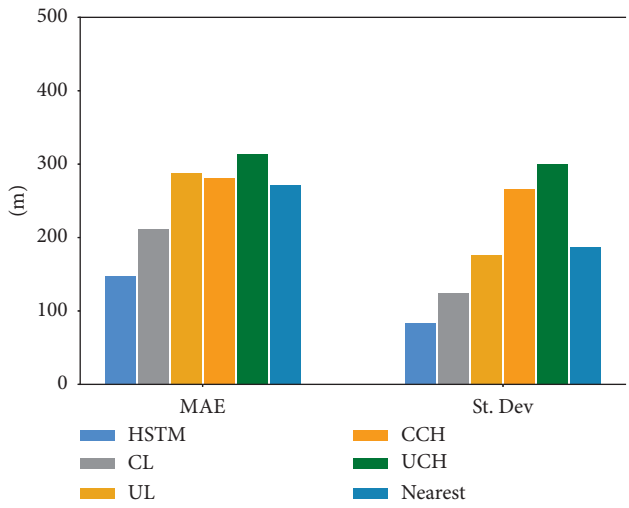

Figure 8: Comparison of five heuristic optimizations.

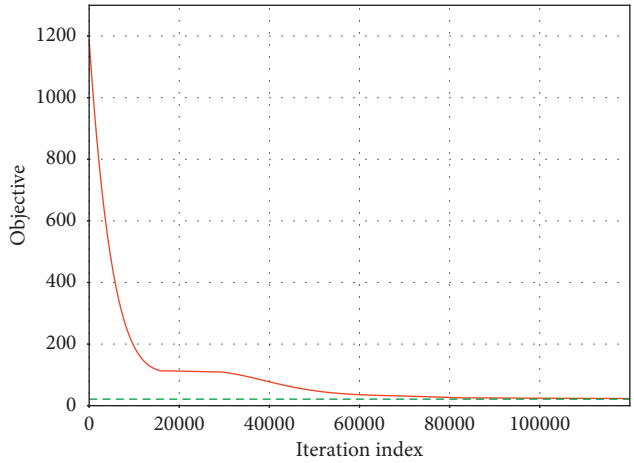

Figure 9: Iterative convergence in compass search optimization.

cell-id trajectories varies from six hundred to two dozen, as shown in Figure 12(a). The errors of estimated trajectory to the GNSS trajectory increase from thirty meters to six hundred meters, as shown in Figure 12(b). The interval of 5 minutes that is close to the average interval of our big cell-id trajectory collection generates an error of 180 m. Once the interval is greater than 10 minutes, our proposed method will not support acceptable trajectory refinement any more.

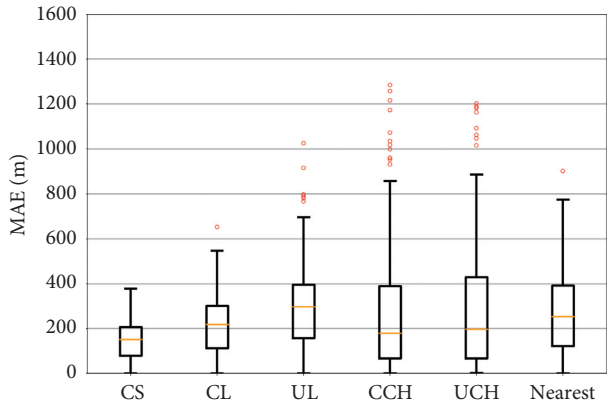

Figure 10: Boxplot of absolute errors of estimated points.

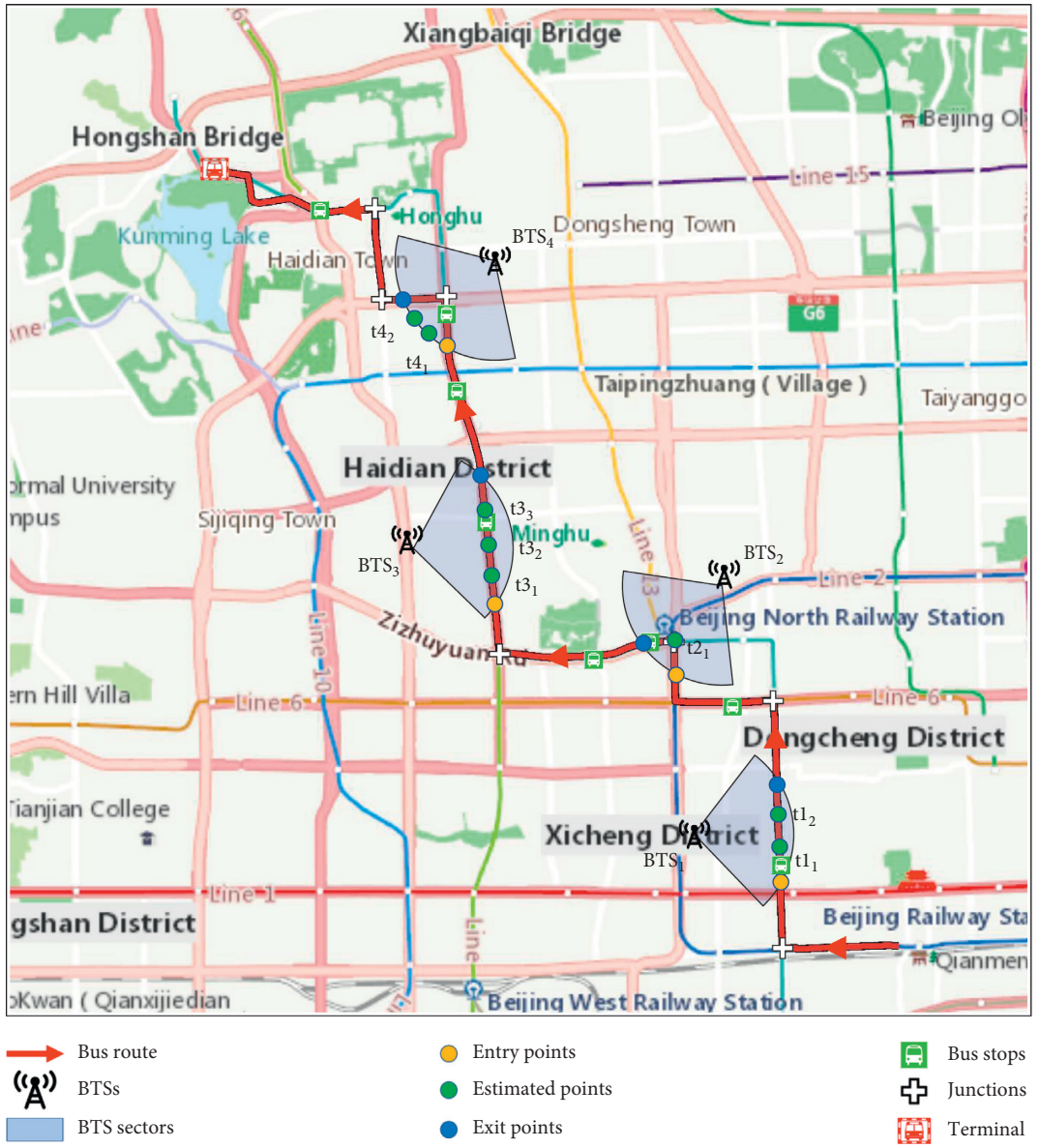

Figure 11: Diagram of BTS coverage and LCSS offset.

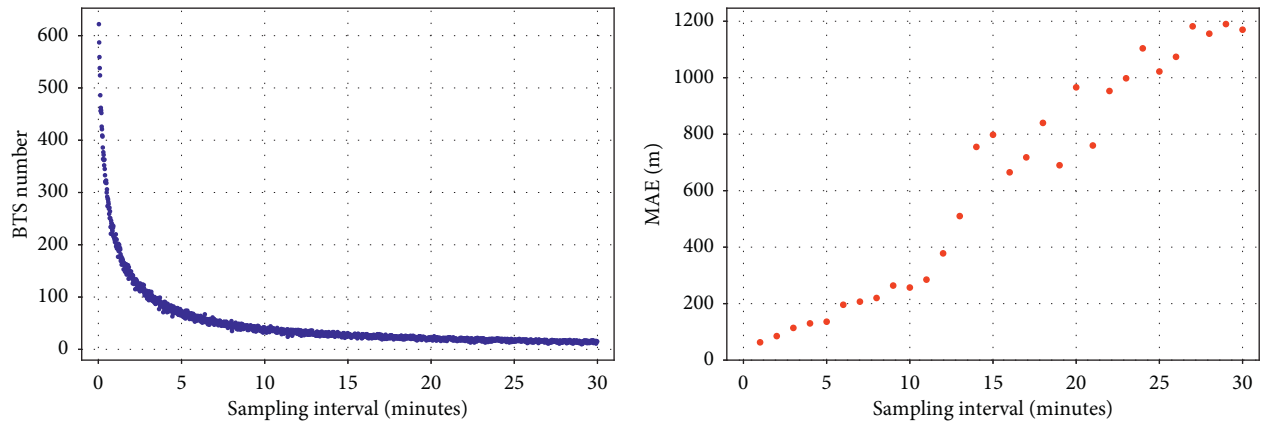

Figure 12: Cell-id trajectories with different sampling intervals. (a) BTS numbers. (b) Errors of estimated anchor points.

6. Conclusion

This work proposed a novel approach to reconstruct the spatiotemporal location of vehicles between sparse updates in a cell-id trajectory. First, an LCSS-based machine learning method was proposed to detect PSV-mode cell-id trajectories from the huge volume of cellular network signaling datasets. Then, a heuristic global optimization method was deployed to estimate the precise locations along bus routes of these detected cell-id trajectories. Evaluated with ground truth datasets, our proposed method has achieved very good performance in both accuracy and robustness in comparison with traditional interpolations.

Our heuristic global optimization with compass search algorithm overcomes the issue of location shifting in LCSS when matching a cell-id trajectory and a bus route. This leads to a set of high-quality anchor points, that is, a spatial position at a road network that is originally corresponding to a cell-id tracking point at a particular time is estimated in the common intersection part of the BTS sector and the road network. The experiment indicates that, by taking advantage of the nature of PSV-mode cell-id trajectories, our approach works well on cellular network signaling datasets with five-minute sampling interval, but the performance decreases sharply after a ten-minute sampling.

Data Availability

The 4G-LTE mobile phone data used to support the findings of this study were supplied by China Mobile Communications Group Co., Ltd., under license and so cannot be made freely available. Requests for access to these data should be made to Kemin, Xianfeng (xfsong@ucas.ac.cn).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key Research and Development Foundation of China (nos. 2017YFB0503702 and 2017YFB0503605), 973 Program (no. 2013CB733402), and National Natural Science Foundation of China (nos. 41601486 and 40771167).

References

[1] K. Liu, S. Gao, and F. Lu, "Identifying spatial interaction patterns of vehicle movements on urban road networks by topic modelling," *Computers, Environment and Urban Systems*, vol. 74, pp. 50–61, 2019.

[2] Y. Xin, *A Stepwise Spatio-Temporal Flow Clustering Method for Discovering Mobility Trends*, p. 1, IEEE Access, New York, NY, USA, 2018.

[3] Y. Xu, "Human mobility and socioeconomic status: analysis of Singapore and boston," *Computers Environment & Urban Systems*, vol. 72, pp. 51–67, 2018.

[4] M. Forghani, F. Karimipour, and C. Claramunt, "From cellular positioning data to trajectories: steps towards a more accurate mobility exploration," *Transportation Research Part C-Emerging Technologies*, vol. 117, 2020.

[5] Z. Smoreda, A. M. Olteanu-Raimond, and T. Couronné, *Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment*, Emerald Group Publishing Limited, Bingley, UK, 2013.

[6] Z. Zhao, "Understanding the bias of call detail records in human mobility research," *International Journal of Geographical Information Science*, vol. 30, pp. 1–25, 2016.

[7] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?" *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 3, pp. 33–44, 2012.

[8] M. Deng, Z. Fan, Q. Liu, and J. Gong, "A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets," *ISPRS International Journal of Geo-Information*, vol. 5, no. 2, p. 13, 2016.

[9] S. Cheng and F. Lu, "A two-step method for missing spatio-temporal data reconstruction," *ISPRS International Journal of Geo-Information*, vol. 6, no. 7, p. 187, 2017.

[10] M. Li, "Reconstruction of human movement trajectories from large-scale low- frequency mobile phone data," *Computers Environment & Urban Systems*, vol. 77, 2019.

[11] X. Gong, Z. Huang, Y. Wang, L. Wu, and Y. Liu, "High-performance spatiotemporal trajectory matching across

heterogeneous data sources," *Future Generation Computer Systems*, vol. 105, pp. 148–161, 2020.

[12] M. Ficek and L. Kencl, "Inter-Call Mobility model: a spatio-temporal refinement of Call Data Records using a Gaussian mixture model," *Proceedings - IEEE INFOCOM*, pp. 469–477, 2012.

[13] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, 2014.

[14] S. Hoteit, *Estimating Real Human Trajectories through Mobile Phone Data*, IEEE, NewYork, NY, USA, 2013.

[15] H. Yu, A. Russell, J. Mulholland, and Z. Huang, "Using cell phone location to assess misclassification errors in air pollution exposure estimation," *Environmental Pollution*, vol. 233, pp. 261–266, 2018.

[16] B. C. Csáji, A. Browet, V. A. Traag et al., "Exploring the mobility of mobile phone users," *Physica A: Statistical Mechanics and Its Applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[17] K. Perera, T. Bhattacharya, L. Kulik et al., "Trajectory inference for mobile devices using connected cell towers," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'15)*, Association for Computing Machinery, New York, NY, USA, 2015.

[18] G. R. Jagadeesh and T. Srikanthan, "Probabilistic map matching of sparse and noisy smartphone location data," in *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems*, Gran Canaria, Spain, 2015.

[19] G. R. Jagadeesh and T. Srikanthan, "Online map-matching of noisy and sparse location data with hidden markov and route choice models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 1–12, 2017.

[20] E. Algizawy, T. Ogawa, and A. El-Mahdy, "Real-time large-scale map matching using mobile phone data," *Acm Transactions on Knowledge Discovery from Data*, vol. 11, no. 4, pp. 1–38, 2017.

[21] Z. L. Xiao, "Lightweight map matching for indoor localisation using conditional random fields," in *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks (Ipsn' 14)*, pp. 131–142, Berlin, Germany, 2014.

[22] G. Chen, A. C. Viana, and C. Sarraute, "Towards an adaptive completion of sparse call detail records for mobility analysis," in *Proceedings of the IEEE International Conference on Pervasive Computing & Communications Workshops*, White Plains, NY, USA, 2017.

[23] L. Bergroth, H. Hakonen, and T. Raita, "A survey of longest common subsequence algorithms," in *Proceedings of the Spire 2000: Seventh International Symposium on String Processing and Information Retrieval*, pp. 39–48, A Curuna, Spain, 2000.

[24] K. Laasonen, "Clustering and prediction of mobile user routes from cellular data," in *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Porto, Portugal, 2005.

[25] G. Yavas, "A data mining approach for location prediction in mobile environments," *Data & Knowledge Engineering*, vol. 54, no. 2, pp. 121–146, 2005.

[26] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories. in Data Engineering," in *Proceedings of the 2002 18th International Conference*, San Jose, CA, USA, 2002.

[27] C. Hermes, "Long-term vehicle motion prediction," in *Proceedings of the Intelligent Vehicles Symposium*, 2010.

[28] H. Alt, *The Computational Geometry of Comparing Shapes*, 2009.

[29] H. Chen, "Exploring human spatio-temporal travel behavior based on cellular network data: a case study of hangzhou, China," *International Journal of Geoinformatics*, vol. 15, no. 3, pp. 1–12, 2019.