MIFF: Human Mobility Extractions with Cellular Signaling Data under Spatio-temporal Uncertainty

YIWEI SONG, Peking University
YUNHUAI LIU*, Peking University, China
WENQING QIU, Peking University, China
ZHOU QIN, Rutgers University, United States
CHANG TAN, iFlytek, China
CAN YANG, Didi, China
DESHENG ZHANG, Rutgers University, United States

Human Mobility Extraction with cellular Signaling Data (SD) is essential for human mobility understanding, epidemic control, and wireless network planning. SD log the detailed interactions between cellphones and cellular towers, but suffer from a spatio-temporal uncertainty problem due to cellular network tower-level load rebalancing (switching users between towers) and cellphone usage activities. To date, most models focus on utilizing better data like RSSI or GPS, do not directly address uncertainty. To address the SD uncertainty issue, we utilize two insights based on (i) individuals' regular mobility patterns and (ii) common co-movement mobility patterns between cellphone users as suggested by fundamental human mobility nature. Accordingly, we design a Multi-Information Fusion Framework (MIFF) to assist in extracting road-level human mobility based on cell-tower level traces. To evaluate the effectiveness of MIFF, we conduct experiments on one-month SD obtained from a cellular service operator, and SD manually collected by handheld mobile devices in two cities in China. Four transportation modes, namely railways, cars, buses, and bikes are evaluated. Experimental results show that with MIFF, our road-level trajectory extraction accuracy can be improved by 5.0% on Point correct matching index and 68.5% on Geographic Error on average.

CCS Concepts: • **Networks** → **Location based services**; • **Human-centered computing** → *Collaborative and social computing*; • **Computing methodologies** → *Model development and analysis.*

Additional Key Words and Phrases: Map Matching, Signaling Data, Homogeneous Traces Search, Multiple Traces Fusion, Regular Pattern Exploration

ACM Reference Format:
Yiwei Song, Yunhuai Liu, Wenqing Qiu, Zhou Qin, Chang Tan, Can Yang, and Desheng Zhang. 2020. MIFF: Human Mobility Extractions with Cellular Signaling Data under Spatio-temporal Uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 159 (December 2020), 19 pages. https://doi.org/10.1145/3432238

*Corresponding author.

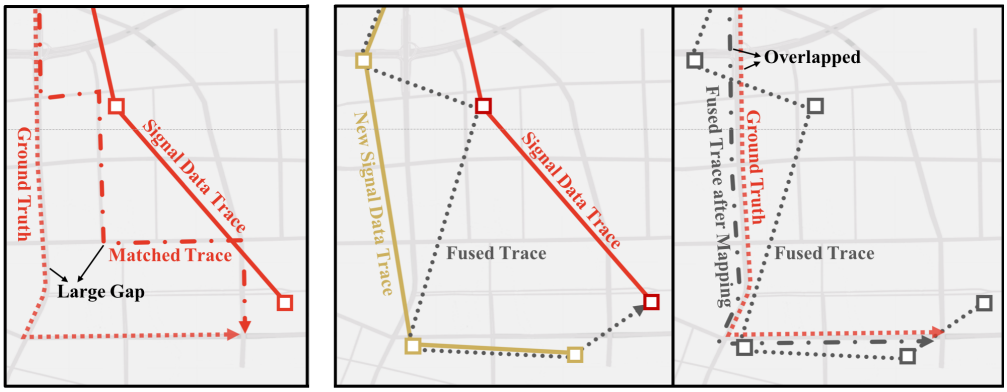
Authors' addresses: Yiwei Song, Peking University, No.5 Yiheyuan Rd. Haidian District, Beijing, China, yiwei.song@pku.edu.cn; Yunhuai Liu, Peking University, No.5 Yiheyuan Rd. Haidian District, Beijing, China, yunhuai.liu@pku.edu.cn; Wenqing Qiu, Peking University, No.5 Yiheyuan Rd. Haidian District, Beijing, China, vikiqiu@pku.edu.cn; Zhou Qin, Rutgers University, New Jersey, United States, zq58@cs.rutgers.edu; Chang Tan, iFlytek, Hefei, China, changtan2@iflytek.com; Can Yang, Didi, Beijing, China, yangcan@didiglobal.com; Desheng Zhang, Rutgers University, New Jersey, United States, desheng.zhang@cs.rutgers.edu.

2474-9567/2020/12-ART159 $15.00
https://doi.org/10.1145/3432238

1 INTRODUCTION

Extracting human traces at a metropolis scale, and mapping them to road networks, are essential for understanding the human mobility[1][9]. There are substantial mobile applications ranging from epidemic control[22], transportation management[1], urban planning[9], to emergence response[18]. For example, COVID-19 is spreading rapidly in many countries and there have been more than 21 million confirmed cases all over the world. It can be greatly helpful for the government to track down people who had been in direct contact with confirmed patients and slow down the spread of the disease outbreak. It can also avoid a catastrophe like a stampede that occurred in Shanghai which resulted in 36 killed casualties and 49 injured[23]. Compare to other spatio-temporal data such as GPS data from a vehicular network or check-in data from a social network app, cellular Signaling Data (SD) are collected to log the signaling level interactions between cellular phones and cellular towers, and SD can cover more residents in the city, which enable more representative human mobility applications[11][6]. For example, human mobility extractions with SD can also benefit cellular network planning by analyzing static and dynamic signaling demands. Especially under the 5G era, cells have a smaller radius and higher density and need exact moving traces during deployments.

Specifically, given the SD of cellphone users at cellular tower level, our goal is to map them to the road network to obtain road level human mobility traces because many human mobility applications are based on road networks. Existing work has made great progress based on fine-grained spatio-temporal data, e.g., for GPS data [26][31]. Some recent works apply deep learning based (DL) algorithms [25][29] to automatically extract trace features from GPS data and have an impressive performance. All these works are, however, proposed for GPS data with relatively low location errors in outdoor environments, and this ensures good performance in these works. These state-of-art methods do not apply to the SD because of the spatio-temporal uncertainty in SD, which will be detailed discussed in Section 2.1. Additionally, DL-based approaches need large-scale labeled training data on each road segment to extracted features. These labeled data for SD traces are scarce and challenging to obtain in practice, which cannot support effective training of DL-based models at the city scale. To alleviate the negative impact of spatio-temporal uncertainty of SD, most of the recent studies on SD[11][17] utilize traditional GPS map matching models by introducing more SD information such as RSSI or RSCP [32], or applying certain prior-knowledge rules such as eliminating the ping-pong effect [15] to enhance the model, under the spatio-temporal uncertainty of SD data. However, none of these studies used any methods to directly address the inherent SD uncertainty but aimed to utilize better data instead.



(a) Mapping without Data Fusion (b) Data Fusion of Homogeneous Traces (c) Mapping with Data Fusion

Fig. 1. Effect of Fusing Homogeneous Traces before Map Matching

SD are the cellular activity records collected for network maintenance, which log the detailed interactions between cellphones and cellular towers. They require no extra human efforts and generate mass cellular tower

level traces. It has been confirmed by the behavior-based research that human mobility is highly repetitive[8]. Thus, there are many **homogeneous traces**, i.e., tower-level traces with the same road-level paths, especially in SD, which have a high penetration rate and a large user scale because most residents have cellphones. To mitigate the SD spatio-temporal uncertainty, our observation is to find homogeneous tower-level traces, fuse them *before*, and then performing map matching. The trace fusion can provide missing information based on multiple traces, and the redundant error can be reduced by smoothing. As a result, our approach is to perform trace fusion before map matching; whereas most of the existing work is to perform map matching before trace fusion[11][17][26][31]. This approach makes our work conceptually different from related work. For example, Fig. 1(a) shows a directly matched trace at the road level from an SD trace without trace fusion. Because of the spatio-temporal uncertainty, the matched trace is far from the ground truth. Fig. 1(b) shows a new signal data trace, which is the homogeneous trace to the original SD traces. If we fuse them, we will have a fused trace shown as the black dashed line in Fig. 1(b). As shown in Fig. 1(c), if we map this fused trace to the road level, the map-matched fused trace has a much smaller gap with ground truth compared to a directly matching without homogeneous traces in Fig. 1(a).

However, the spatio-temporal uncertainty of SD makes it challenging for us to find homogeneous traces and perform a fusion on these traces. A most straightforward intuition is to first obtain matched road traces by matching the cell tower level traces to the road network and use the road traces as an intermediary to find the homogeneous trace. It is not feasible because directly matching the cell tower level trace is not accurate at all, which makes it hard to find homogeneous traces. Our key idea is to find homogeneous traces by (a) regular mobility patterns of the same users between different days (e.g., daily commuting), and (b) the co-movement phenomenon of different users in the same time period. For example, many people travel with public transportation, e.g., subways and buses, which results in mass partial similar traces [31]; even in a taxi, the passengers and the driver's co-movement produces at least two similar traces[1]; finally, many people have regular mobility patterns, e.g., the way they go to work/school[8] from the same residential areas. Here we propose a Dynamic Time Warping (DTW) based model to align homogeneous traces before fusion.

In this paper, we design and evaluate a new human trace extracting model MIFF based on SD, which selects co-moving traces and individuals' regular patterns, and then fuses these homogeneous traces to obtain a denser trace before map matching to alleviate the spatio-temporal uncertainty. *We discuss the privacy and ethic issues in the Discussion section.* The contributions of this paper are as follows:

- Conceptually, we conduct a large-scale empirical study to investigate the human mobility regularity and similarities to extract road-level traces based on cell tower SD. In particular, we identify a primary challenge of finding cell tower trace homogeneous traces to perform trace fusion before map matching to reduce the negative impact of spatio-temporal uncertainty of SD. We further addressed this challenge by investigating (a) the regular mobility patterns of the same users between different days, and (b) the co-movement phenomenon of different users.
- We design a multi-information fusion framework(MIFF) for mobility extraction based on cellular signaling data. Based on the key characteristics of cellular signaling data, we explore a fundamentally different strategy to fusion trace data first and then map them to a road network, while the straightforward idea is to perform the mapping first and then fuse the road level trajectory.
- To evaluate the overall performance of MIFF, we explore two large-scale complementary signalling datasets in two different cities (i.e., Chinese city HeFei and Shanghai) with four different transportation modality, and the results show that the data fusion in MIFF can effectively increase by 5.0% on Point correct matching index and 68.5% on Geographic Error compared to the state-of-the-art models.

2 BACKGROUND AND MOTIVATION

In this section, we (i) describe the characteristics of SD, focusing on the spatio-temporal uncertainty and new opportunities that SD offer to us; (ii) identify challenges in finding homogeneous traces when applying SD to build the mobility model; (iii) illustrate our observations on the regularity of human mobility and two design insights.

2.1 Characteristics of Signaling Data

SD are generated from cellular service providers when a cellphone interacts with the associated cell towers. In general, SD have the data format with the time-stamp, cell ID, user ID (anonymous ID), and signaling types. SD are the superset of the Call Detail Records (CDR) with the difference that CDR only record information during phone calls for billing purposes, while SD are much denser due to various recording for different cellular activities at signaling levels, e.g., attach or detach from a cellular tower.

Compared with cellular CDR, which may not have any information when users have no call or text activities, periodic SD keep offering updated states and checking for the existence of the cell phone with paging, attaching and detaching, and thus SD show a relatively stronger temporal continuity than CDR. Compared with GPS data, SD have three main merits in term of transparency, user coverage, and user behavior coherence. (i) SD are more transparent because they are automatically generated in the cellular system. The cellar service operators can log all the records without installing any new devices or consuming any extra energy. (ii) SD usually keep a higher user coverage rate. For example, in China, there are three mobile phone operators, China Mobile, China Unicom, and China Telecom. Their market share are around 62.4%, 20% and 17.6%[24]. The users of each operator are obtained by relatively unbiased sampling from all residents in the city. (iii) The underlying user behavior capturing of SD is more continuous. SD are always recorded as long as people turn on their cellphones. Even in the standby mode, some SD service type data (e.g., paging) are still recorded when users travel for a long distance or standby for a certain while, e.g., 20 mins. In contrast, GPS data are often turned off for privacy and energy consumption concerns while not in use.

Although SD have the advantages mentioned above, the spatio-temporal uncertainty makes us unable to straightforwardly apply the state-of-art model mentioned in Section 1. In particular, the spatial uncertainty comes from the cellular tower switching strategy as well as load balance strategy for mobile users, making it very different from GPS data with noise obeying a Normal distribution [31]. The temporal uncertainty comes from the user's irregular usage frequency, e.g., phone calls, surfing the Internet, or recent APP-level data usage. These two defects are discussed in detail below

- **Spatial Uncertainty**. Generally, the cell radius of a single cellular tower is about 300 meters to 1000 meters [11]. As the cellular tower selection strategy is not based on the proportion of the signal distance [20], a cell phone may connect to a cellular tower 600 meters away, even if it is 300 meters away from the nearest cellular tower nearby. Besides, since the load balance strategy may randomly distribute cellphones to cellular towers within the range [11], the true location of cellphones can appear anywhere within the coverage of the tower. So the recorded cellular tower in SD is normally not the one closest to the actual location of cellphones, which makes the distance-based localization algorithms fail to work.
- **Temporal Uncertainty**. SD are temporally sparse and irregular. Although a standby cellphone logs periodic SD to inform the cell tower of its existence and the current state, these SD are still too sparse to be mapped into accurate road segments like GPS data. To demonstrate the overall statistical characteristic, Fig. 2 depicts the probability density plot of SD's temporal interval. Compared with the fixed time interval of GPS data (e.g., 10 seconds) [17], the average time interval of SD is 131.47 seconds, and there are 24.1% users with an average time interval larger than 3 minutes and 9.13% users larger than 5 minutes. Further, Fig. 3 illustrates a user's SD trace with an average time interval of 3.97 minutes. The left figure shows the

whole SD trace. The right figure shows the amplified partial trace between point B and point C. The green and red dashed line are only two possible road paths travel from B to C. It is challenging to distinguish the exact road trajectory without extra information.

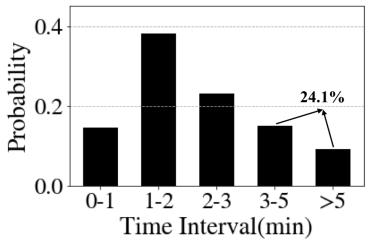


Fig. 2. SD time interval probability plot

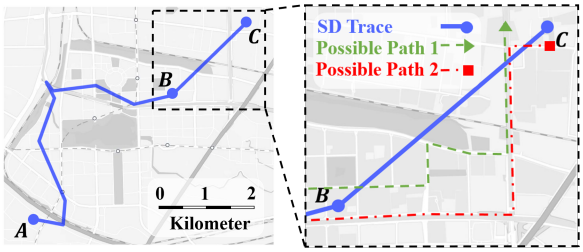
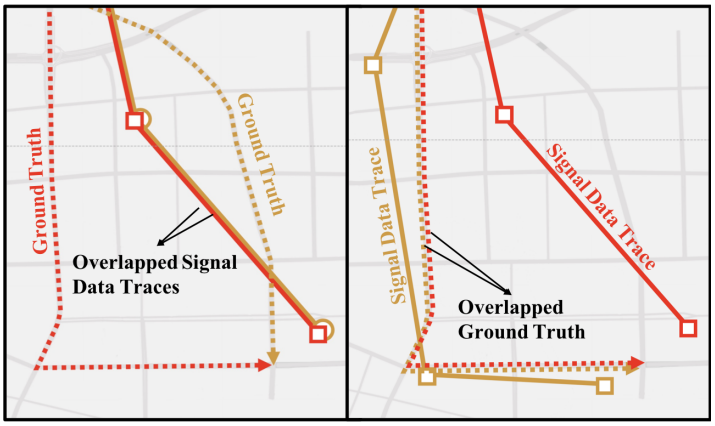


Fig. 3. One SD trace with long time interval.

2.2 Design Challenges

We have illustrated that the tower-level SD traces with insufficient information is challenging to map to the actual road level trajectories due to spatio-temporal uncertainty. Intending to obtain more accurate matching results, we focus on reducing the negative impact of uncertainty by fusing homogeneous traces. Despite many studies looking for homogeneous traces, or more generally similar traces, this task keeps challenging. Specifically, for the spatio-temporal uncertainty problem, finding homogeneous traces before map matching is the main challenge of our work.



(a) Same SD traces with very different Ground Truth

(b) Very different SD traces with the same Ground Truth

Fig. 4. Illustration of Homogeneous Traces

On the one hand, it is intuitive to find homogeneous traces according to ground truth in terms of similar traces. Even if similar traces are found, a large number of which may not be homogeneous traces. For example, in Fig. 4(a), two same SD traces have different ground truth. While in Fig. 4(b), two dissimilar SD traces have the same ground truth. This is mainly due to the spatial uncertainty of SD. Hence, it is difficult to distinguish whether a group of similar SD traces belong to the same road-level trajectories without the assistant of ground truth. Even worse, as we mentioned above, the shortcoming of insufficient ground truth is something we cannot make up for. On the other hand, the road level traces obtained through map matching is very helpful to find similar traces. For example, the traces composed of GPS data will be matched with the map first, and the matched road paths are used to calculate the similarity to obtain the similarity relative to the GPS traces. However, the spatio-temporal

uncertainty of SD data determines the unreliability of its map matching results. If the map matching result is not credible, there is no way to find similar traces.

2.3 Design Insights

Here we define **Homogeneous Traces** as a set of cell-tower level traces with the same road network level trajectories. Given a group of homogeneous traces, it is a good option to first fuse the traces into a more informative trace. Then we can map this fused trace to a road-level trace by map matching models. Here is an example. Fig. 6 adds two homogeneous traces of the blue trace, i.e., the red and green dash ones in Fig. 3. The red and green traces provide sufficient information between point B to point C. Thus the real road level trajectories of the blue trace are probably Possible Path 2 shown in the right of Fig. 3. Considering the challenge of finding homogeneous traces, we must enhance the understanding of the characteristics of SD. The great user coverage and high penetration rate of SD provide us access to investigate the mobility and movement patterns of cellular users. In this paper, we argue that thanks to individual regularity and user behavior coherence, we can improve our homogeneous trace identification based on two design insights as follows.

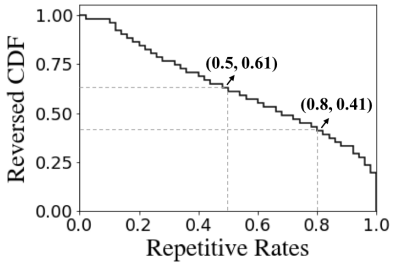


Fig. 5. Reversed CDF plot of people's repetitive rate in GPS data.

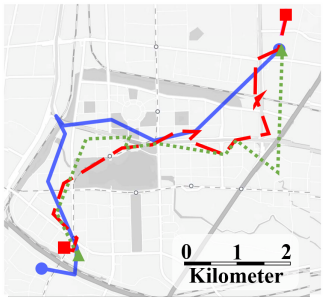


Fig. 6. Homogeneous Traces of Same Users

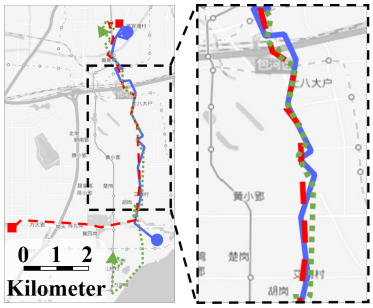


Fig. 7. Different users' traces with time scale constraint.

Insight 1: Regular Pattern of the Same Users. People usually have regular patterns based on their social habitats[8]. For example, they travel in the same path to go to work, go to school, go shopping, etc. Thus the similar traces in the history traces of one person are likely homogeneous. We conduct an empirical study on human regular patterns to show the repetitive rate of each person. The data collect 24,432 private vehicles' GPS data from 04/2016 to 06/2016 in Shanghai, covering 19,724 roads with 7.9 million GPS records every day. We label a trace as a repetitive trace if there are similar traces in that person's history. The repetitive rate is the ratio of repetitive traces in the whole traces for each person. The cumulative distribution function (CCDF) plot of repetitive rates is shown in Fig. 5. It illustrates that more than 30% of users have a repetitive rate higher than 80%; nearly 60% of users have a repetitive rate higher than 40%, which indicates that repetitive traces are common for most users. Additionally, we conduct a set of case studies with SD from users with highly regular patterns. Fig. 6 is one example of them and shows three traces from one person on three different days. These traces are all recorded from 8:00 AM to 9:00 AM and their travel times are respectively 43.7 minutes, 31.6 minutes, and 33.8 minutes.

Insight 2: Co-movement between Different Users. We explore the co-movement phenomenon in large cities. Many people travel by public transportation (e.g., buses and subways), meaning that these people have partial homogeneous traces with each other [13]. Further, people travel with private vehicles are also possible co-riding with others on the same road segment [16]. For example, three different users' traces are given in Fig. 7 where they shared a very long path due to their co-movement activities.

To sum up, we utilize two insights in this paper as key opportunities to address spatio-temporal uncertainty of SD: a person is more likely to travel on the same road-network level path if two raw tower-level traces belong to him/her are similar; people are more likely to travel in the same road-network level path if their raw tower-level traces are similar to strict temporal constraints. As a result, how to measure trace similarity at tower level and resultant trace fusion are the key in our design, which will be given as follows.

3 SYSTEM DESIGN

The architecture of our system is shown in Fig. 8. It is composed of five major components: input SD data, trajectory segmentation, preprocessing machine (PM), multi-information fusion framework (MIFF) and map matching. As follows, we will depict the whole system details shown as below, not only the four major components. (a) Data input and notions will be given to help understanding and applying our system; (b) Trajectory segmentation is to split each users' SD into several moving trajectories; (c) Preprocessing machine is to obtain preprocessed single traces and stay points. (d) MIFF is a novelty framework fuse multiple information to obtain common stay points, regular patterns, co-moving traces and fused traces. We further split MIFF into two parts to illustrate: homogeneous traces search and data fusion.

Note that in this paper, we focus on the homogeneous traces search and data fusion. Any existing or new map matching algorithms can be applied to our framework. The map matching part is not our main contribution. We already implemented several existing map matching algorithms, and the implementation details are shown in Section 4. The two-stage smoothing method is mentioned in Section 3.2.2.

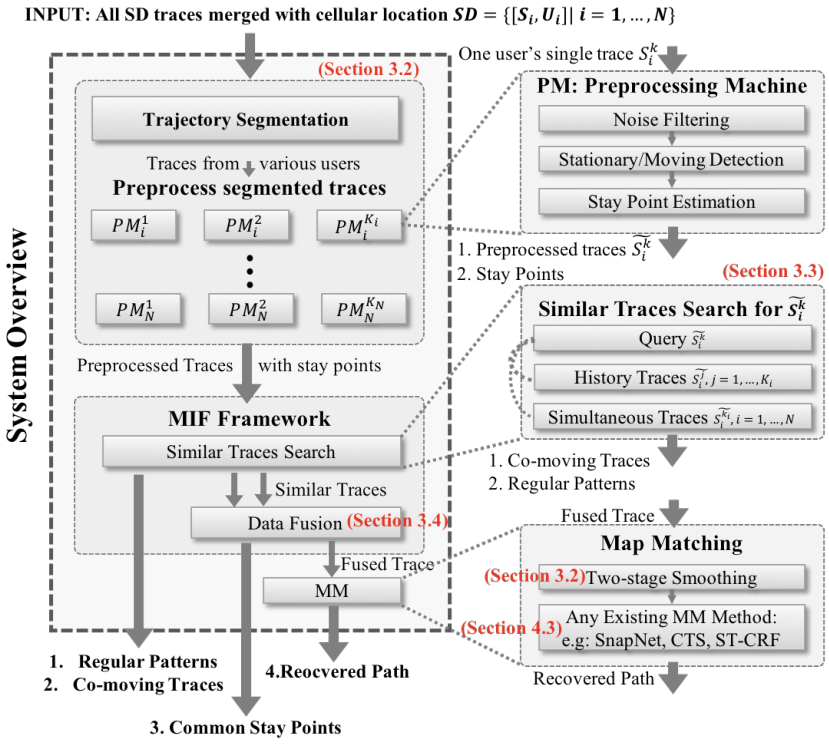


Fig. 8. System Overview

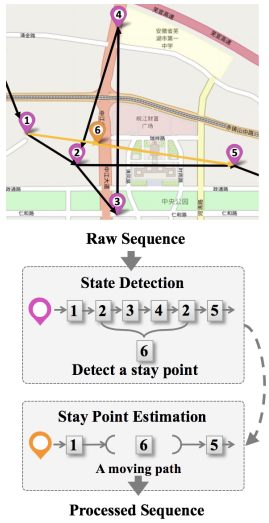


Fig. 9. Cell tower level State Detection and Stay Point Estimation Illustration. Black line is the raw sequence, and the yellow line is the sequence after stay point estimation.

3.1 Data Input

There are three main data involved in our system: SD Trajectories Logs, Cellular Locations, and Road Network. SD Trajectories Logs consist of timestamp, user ID (with desensitization), and cell ID. Cellular Locations consist of cell ID, longitude, latitude. We do not use any sensor data and cellular attributes as the input of our system, such as nearby 6 cell towers at one observation and their signal intensity in CTS[10], which can be used for trilateration positioning. Here are some definitions and notions which will be used in MIFF.

Definition 1. ***SD Data***. *SD data contain N users' cell-tower trace shown as* $SD = [S_i, U_i]|i = 1, ..., N$ *where* U_i *is the* i_{th} *user, and* S_i *is the cell-tower trace set of* U_i.

Definition 2. ***A cell-tower trace***. *Given the trace set* S_i, *the* k*th trace* $S_i^k = \{s_{i,1}^k, s_{i,2}^k, \ldots, s_{i,|S_i^k|}^k\}$ *is one temporally ordered sequence split from the* i_{th} *user's trace set* S_i. *Each observation* $s_{i,t}^k$ *is a spatio-temporal point with dimension <longitude, latitude, timestamp, features>.* $|S_i^k|$ *is length of the trace* S_i^k.

3.2 Trajectory Segmentation and Preprocessing Machine

We take the daily SD records of a user as input and convert them into low-noise traces with timestamps, longitude, and latitude. Then we use stationary moving detection to identify stay points and moving traces.

3.2.1 Trace Segmentation. Unlike many GPS datasets with extracted traces as basic elements[28, 30], the collection of SD independent of the user's movement patterns. We need to extract the trajectory from the SD data generated when the user moves. Given one user's whole day SD, we need to segment the SD records into several reasonable traces with starting points and destinations. We segment the records by stay point detection to obtain traces based on our empirical validation of a state-of-art algorithm [31]. The maximum time interval is 20 minutes. Besides, travel distance and travel points are required not smaller than 1km and 5 points to ensure that the trace has a meaningful length given the observation from[31].

3.2.2 Noise Filtering. During SD data logging, many kinds of noise appear randomly. A ping pong effect appears when the phones moving in a slow speed [17]. Concretely, in a period of time, the cell phone does not maintain a connection with a single cellular station. In fact, it may connect to several connectable cellular stations randomly. The greater the density of cell tower, the greater the randomness. Additionally, some outliers such as recording errors are also inevitable, which leads to several sudden shifting, deviating more than 5km or even 10km. In this paper, we try to eliminate the effect of these errors by the following two methods. **(1) Outlier filtering** is applied to the raw sequence S_i^k splitted by trajectory segmentation. We eliminate outliers by using some speed-based noise filtering and direction-based noise filtering algorithms shown in CTS[10]. **(2) Two-stage smoothing** is used on the fused trace shown in the Fig 8 to flat the sensor error and ping pong effects. The first stage moving average smooths the large sensor error with a low moving window. The second stage moving average is conducted after time interpolation, which customizes the time interval between two adjacent points. In that way, the second stage moving average with a low moving window can flat the raw trace into a smooth line.

3.2.3 State Detection and Stay Point Estimation. State detection is the stationary/moving detection shown in the Fig. 8. We detect the stationary state on two levels. (1) Cell tower level state detection is first to be applied. We found that when a user stays at a point for a short time, there usually is some ping-pong effect. Fig. 9 is an example where the black line in the figure is the raw trace. It appears a circle subsequence $(2 \rightarrow 3 \rightarrow 4 \rightarrow 2)$, which is obviously a stay point. In this example, this stay point is a wait for a traffic light. (2) Geo-level state detection applies the method proposed in CTS[10].

3.3 Homogeneous Traces Search

With the data preprocessing, the records of users are segmented to a large number of individual traces. The goal of homogeneous traces search is to find groups of SD traces with a high probability of belonging to the same road-level traces. The candidate trace can either be a full trace or a partial trace segment during a trip. Note that there are two kinds of homogeneous traces: those belong to the same person at different time, and those belong to different persons at the same time. As follows, we first give the definition and notations that will be used in our work, and then a DTW-based mobility similarity measurement algorithm will be introduced. Finally, we show how to identify the two kinds of homogeneous traces.

3.3.1 Mobility Similarity Measurement. Due to the nature of SD, the homogeneous traces in this paper mean that two tower-level traces are potentially belonging to the same road-level trace. To quantitatively measure this, we design a mobility similarity measurement algorithm based on Dynamic Time Wrapping (DTW). As DTW assumes the same sampling rate, we fill the original traces with the identical time interval by linear interpolation to obtain new traces. The DTW distance between two raw traces is shown below. Here S_*^1 and S_*^2 are two SD traces for arbitrary users. $s_{*,t}^k$ means the t^{th} records of S_*^k. $Rest(*)$ refers to the part of S_*^k left after $s_{*,t}^k$ is removed.

$$\begin{aligned} DTW(S_*^1, S_*^2) &= D(s_{*,1}^1, s_{*,1}^2) + \\ &\min \begin{cases} DTW(S_*^1, Rest(S_*^2)), & S_*^1 \text{stutter} \\ DTW(Rest(S_*^1), S_*^2), & S_*^2 \text{stutter} \\ DTW(Rest(S_*^1), Rest(S_*^2)). & \text{no stutter} \end{cases} \end{aligned} \tag{1}$$

As follows, D function is to calculate the geographic distance between two positions s_i and s_j, where Δlon and Δlat is the difference between s_i and s_j in terms of the longitude and latitude. $D(s_i, s_j)$ is given as

$$2Rsin^2(\Delta lat) + cos(s_i.lat)cos(s_j.lat)sin^2(\Delta lon/2) \tag{2}$$

DTW reconstructs the traces with index identity $\{s_{*,w_1}^k, s_{*,w_2}^k, \ldots, s_{*,w_t}^k, \ldots, s_{*,w_T}^k\}$, where w_t is the index of the sequence and $1 \le w_t \le w_{t+1} \le T$. T is the trace length $|S_*^k|$. Hereon, the DTW distance between traces S^1 and S^2 is defined as

$$DTW(S_*^1, S_*^2) = \frac{1}{T} \sum_{t=1}^{T} D(s_{*,w_t}^1, s_{*,w_t}^2) \tag{3}$$

The time interpolation may introduce extra errors, hence a confidence of each point s_{*,w_t}^k is proposed according to whether it's a raw point or an interpolated point. The raw point have a 100% confidence. And for the interpolated point, the closer to the raw point, the higher is its confidence. Let t be the minimum time from the interpolated point to the raw points. σ infers the distance between the interpolated and the corresponding raw point. The confidence is defined as below.

$$FC(s_{*,w_t}^k) = \begin{cases} 1, & s_{*,w_t}^k \text{ is a raw point} \\ 2\int_0^t \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \mathrm{d}x, & s_{*,w_t}^k \text{ is an interpolation} \end{cases} \tag{4}$$

Therefore, the confidence DTW distance (CDTW) with confidence as weights is expressed as follows.

$$\begin{aligned} CDTW(S_*^1, S_*^2) &= \frac{1}{Z(S_*^1, S_*^2)} \sum_{t=1}^{T} W(s_{*,w_t}^1, s_{*,w_t}^2) D(s_{*,w_t}^1, s_{*,w_t}^2) \\ W(s_{*,w_t}^1, s_{*,w_t}^2) &= FC(s_{*,w_t}^1) FC(s_{*,w_t}^2) \end{aligned} \tag{5}$$

where $W(s_{*,w_t}^1, s_{*,w_t}^2)$ is the weight of $D(s_{*,w_t}^1, s_{*,w_t}^2)$; $Z(S_*^1, S_*^2)$ is the weights sum.

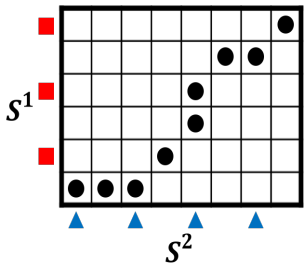


Fig. 10. Alignment Relationship between S^1 and S^2.

3.3.2 Regular Pattern Search of the Same User. Many people have habitat-based regular mobility patterns, e.g., going to work, going to school, or go shopping. Under this intuition, it can't be denied that similar traces from the same user are more likely to be homogeneous traces. Thus, we search the history traces of each user and compute their CDTW in pairs according to equation (6). A maximum CDTW threshold is set to judge whether two traces are homogeneous. Multiple traces are homogeneous if all the traces pairs are homogeneous.

3.3.3 Co-movement Pattern Search among Different Users. For traces that belong to different users, the similarity is hard to measure. But it is common sense that people usually travel with a group of strangers such as on the same bus or train. In that case, searching similar traces that happened at the same time for map matching is effective. Sometimes, people apart from a fixed distance (e.g., people are on the same subway or the same road) can also have a similar trace. Considering this situation, a fixed time translation is allowed but with a time penalty $\triangle\ ts_t$ between s^1_{*,w_t} and s^2_{*,w_t}. In the experiment, we make the coefficient α equal to 0.5. Finally the confidence of time consistent DTW distance $CTDTW(S^1_*, S^2_*)$ is defined as below.

$$\begin{aligned} CTDTW(S^1_*, S^2_*) &= \frac{1}{Z(S^1_*, S^2_*)} \sum_{t=1}^{T} \left(ctdtw(s^1_{*,w_t}, s^2_{*,w_t})\right) \\ ctdtw(s^1_{*,w_t}, s^2_{*,w_t}) &= W(s^1_{*,w_t}, s^2_{*,w_t}) D(s^1_{*,w_t}, s^2_{*,w_t}) + \alpha \triangle\ ts_t \end{aligned} \tag{6}$$

3.3.4 Partial Traces Search. With the above homogeneous traces search method, some partially homogeneous traces can not be found. We argue that two people have the same traces as unusual and it is more common that people only move together for a while. Therefore, partial traces search is an important issue. During computing CTDTW, we first the compute the confidence temporal constrained distance between s^1_{*,w_t} and s^2_{*,w_t}, and get the sequence $\left\{ctdtw(s^1_{*,w_t}, s^2_{*,w_t}) | t = 0, ..., T\right\}$. The CTDTW distance is the average of the above sequence. Thus partial homogeneous traces means there exists a start index w_s and an end index w_e, which makes the average of $\left\{ctdtw(s^1_{*,w_t}, s^2_{*,w_t}) | t = s, ..., e\right\}$ is smaller than the threshold.

3.4 Mobility Data Fusion

Given several homogeneous SD traces, we need to fuse them into one trace for map-matching. In the next, we will introduce how to fuse two traces and how to fuse an arbitrary number of traces one by one.

3.4.1 Fusion with Two Traces. For two similar traces S^1 and S^2, we design an alignment relationship function w. With the alignment relationship, the reasonable inner order of the two raw traces will be found to produce a new fused trace.

For example, the alignment relationship between S^1 and S^2 is shown in Fig. 10. Red squares $\{s^1_2, s^1_4, s^1_6\}$ and Blue triangles $\{s^2_1, s^2_3, s^2_5, s^2_7\}$ are the raw tower level points. The fused trace is given by black dots, i.e., $\{s^2_1, s^2_3, s^1_2, s^2_5, s^1_4, s^2_7, s^1_6\}$. It takes O($n$) time to index two traces with DTW using FastDTW where n=$|S^1| + |S^2|$.

Algorithm 1 Multiple Data Fusion

Require:
m similar traces after interpolating with a short fixed time interval, $S^1, \ldots, S^m$.
Ensure:
One trace after fusion for map matching, S^*.
1: Initialize: $S^* = S^1$
2: **for** each S^i in $\{S^2, \ldots, S^m\}$ **do**
3: Perform two traces data fusion on S^* and S^i to get a new S^*.
4: **end for**
5: **return** S^*

3.4.2 Fusion with Multiple Traces. For m similar traces, the time complexity of multiple DTW is about $O(n^{m-1})$. We propose an $O(nm)$ method to re-index m similar traces and fuse them into one single trace. The algorithm is shown in Algorithm 1.

4 EXPERIMENTS

In this section, we conduct a data-driven evaluation in Hefei SD and a field study in Shanghai SD. Hefei SD, thereafter **CTCC Hefei Data**, comes from the mobile phone operator CTCC with massive users in Hefei. Considering the scarcity of labeled data in Hefei (with which we have the ground truth), we manually collect Shanghai SD, thereafter **Shanghai Data**, by our team with handled devices. These two SD sets contain different transportation modes such as train, car, and bus. We show the performance under various transportation modes, temporal frequency, spatial shifting.

4.1 Dataset Description and Preparation

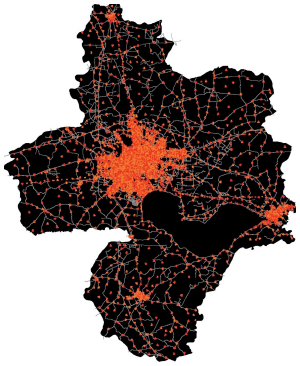


Fig. 11. Cells in HF

Fig. 12. Road Network in HF

Fig. 13. Road Network in SH

4.1.1 CTCC Hefei Data. CTCC Hefei Data contain information about cell tower ID, timestamp, and anonymous user ID. They collect 3.6 million cellphone users' SD for 30 days from 2017/06/01 to 2017/06/30, with 256.9 million records per day. On average, every user generates 71.4 records per day. The 23,704 cell towers of Hefei are displayed in Fig. 11, where a red dot represents a cell tower. Road network data are essential when map traces onto the roads. We collect the road networks of Hefei (similarly Shanghai) in OpenStreetMaps. The Hefei road network in Fig. 12 contains 5,901 road links and 32,708 road segments. It has a vertical length of 84.45km and a horizontal length of 96.48km.

The key drawback of CTCC Hefei Data is that it does not contain the real-time actual positions of users on road networks, i.e., ground truth. Without this, it is hard to evaluate our model. However, a person's trajectory may be recorded by both cellular activities and GPS devices. Thus, we matched the GPS traces from 150 taxis' orders with CTCC Hefei Data for ground truth. They are collected during 2017/06/01-2017/06/02 in Hefei by DiDi [5], a Chinese ride-sharing company. The average time interval of Didi' GPS data is 3.55s, and the average travel time of each order is about 58.8 minutes. 13 groups of homogeneous traces were matched via the homogeneous traces search detailed in 3.3. The average time interval of matched CTCC Hefei Data is 85.89s, and the average Geographic distance from SD to ground truth is 354.8m.

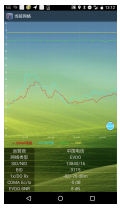


Fig. 14. Interface of SD Collection APP

Table 1. Summary of Shanghai SD

Transportation modes	Time Intervals (s)	Speed(km/h)	# Trace Groups
Train	27.91 ± 0.1525	222.89	70
Car	5.0 ± 0.524	32.67	10
Bus	9.6 ± 0.697	18.03	7
Bike	14.46 ± 4.466	11.09	4

4.1.2 Shanghai Signaling Data. Except for real-world SD, we'd like to study various performance on different transportation modes. As shown in Table 1, Shanghai SD is *manually collected by our team* with handled devices (Huawei Nexus 6P) by the APP developed by CTCC. The interface of the APP is shown in Fig. 14, which shows the information about Section ID, Node ID, and Block ID (identifying the cell towers). Additionally, we utilize Google Map to get real-time GPS data as the actual positions. We ask four volunteers to carry 4 devices to especially collect the SD along with GPS data. Finally, we get Shanghai SD for four kinds of transportation modes (train, car, bus, bike). There is no subway because GPS Data is hard to obtain underground. The average temporal intervals, moving speeds, and the number of homogeneous trace groups are shown in Table. 1. The road network of Shanghai in Fig. 13 contains 46,322 road links and 239,977 road segments. It has a vertical length of 138.7km and a horizontal length of 121.62km.

4.2 Evaluation Metrics

We employ two main metrics, i.e., F1 score (calculated by precision and recall) and geographic errors, to evaluate the accuracy of our MIFF system compared with other methods.

Road-segment-level metrics: To evaluate whether our predicted path match the real path, we use *point correct matching precision (PCMP)*[14] and *point correct matching recall (PCMR)* calculated by the following equation:

$$PCMP = \frac{correct\ matching\ points}{number\ of\ points\ to\ be\ matched} \times 100\% \tag{7}$$

$$PCMR = \frac{correct\ matching\ points}{number\ of\ points\ of\ ground\ truth} \times 100\% \tag{8}$$

Point correct matching index (PCMI) is a comprehensive road-segment-level metric, which combines PCMP and PCMR by the following equation:

$$PCMI = \frac{2(PCMP \times PCMR)}{PCMP + PCMR} \tag{9}$$

Distance-based metrics: PCMP merely consider the number of road segments aligned. For unaligned road segments, we need a metric to evaluate how far these unmatched road segments are. *Geographic Error (GE)* is the average minimum distance between the raw sequence S and the real path P.

$$Geographic\ Error = \frac{1}{|S|}\sum_{s}\min_{p\in P} D(s,p) \tag{10}$$

4.3 Implementation of MIFF and baselines

After homogeneous traces search, we divide cell tower level traces into groups. In our system, all the exists map matching algorithms are suitable for our model. Here, we implement two baseline models, HMM based model(**HMM**) [21] and cellular map matching model(**CTS**) [10].In the experiment, the input of these two models is single cell tower trace without fusion.

- **HMM [21]**: In HMM-based model, the main task is to optimize a Markov chain by two main parts need to heuristically design is the observation score $P(s_n|p_n) = \frac{1}{\sqrt{2\pi}\sigma_z}e^{-0.5\left(\frac{d^s(s_n,p_n)}{\sigma_z}\right)^2}$ and the transition score $P(p_{n-1}|p_n) = \frac{1}{\beta}e^{\frac{-d_t^\beta}{\beta}}$ where $d_t^\beta = |d^s(p_{n-1},p_n) - d^p(p_{n-1},p_n)|$. Here, $d^s(s_n,p_n)$ means the geographic distance between the observation point and the actual location; $d^s(p_{n-1},p_n)$ means the geographic distance between the observation point at timestamp $n-1$ to n; $d^p(p_{n-1},p_n)$ means the minimum route distance from actual location at timestamp $n-1$ to timestamp n. As the optimal parameter in HMM is based on the characteristics of the given data, we estimate the new σ_z and β separately for signaling data. In theory, σ_z is estimated with the median absolute deviation (MAD) as $\hat{\sigma}_z = 1.4826 \times \text{median}_t(d^s(s_t,p_t))$. β is estimated by a robust estimator proposed by Gather and Schultze [7] as $\hat{\beta} = \frac{1}{ln(2)}\text{median}_t\left(|d^s(p_{n-1},p_n) - d^p(p_{n-1},p_n|\right)$.
- **CTS [10]**: CTS is an SD data based enhancement algorithm of the Hidden Markov Model (HMM). In CTS, it uses many cellular underlying data such as signal travel time and cellular basic data such as Antenna direction and radiation angle of the sector. But these data are not provided in our data. Without signal travel time, we can not use the trilateration positioning algorithm, which can reduce the geographic error to a large extent. Without the Antenna direction and radiation angle of the cellular sector, we can not filter the candidate roads by the signal direction. Nevertheless, CTS still improves the HMM model concerning state detection, noise filtering, punishment on detour error, and u-turn errors.

To verify the influence of spatio-temporal uncertainty on trajectory fusion and map matching, we also propose another two baseline models, which are **HMM_RF** and **CTS_RF**, i.e., Road-level trace Fusion. In contrast, MIFF used cellular-tower level trace fusions. What we want to compare are (1) using road level trace fusions to find homogeneous traces and (2) using tower level trace fusions to find homogeneous traces. More specifically, we input each cell tower level trace into a basic map matching model to get a matched road level trace separately and then perform road level trace fusion on each homogeneous trace group. As for road level trace fusion, we extract the most frequent road segment in each time slot (1 minute) for each group of traces. If the adjacent road segments are not connected, the shortest path algorithm is adopted to search for the path connecting the two road segments. Finally, for each group of homogeneous traces, we get a fused road trace.

Further, we implement MIFF by using **HMM** and **CTS** separately as existing map matching module shown in Fig. 8, which are **HMM_MIFF** and **CTS_MIFF**. In these two models, we first fuse each group of homogeneous traces to obtain a fused trace and then perform map matching with the fused trace in the two basic models respectively. The method of cell tower trace fusion is described in Section 3.4 as an important module of MIFF.

4.4 Experiment Results

Here we evaluate our overall performance, including preprocessing machine, homogeneous traces search, and path recovery after data fusion. Then to illustrate the detailed improvement, we show the different performance on various transportation modes, various insights of identifying homogeneous traces, and temporal frequency.

4.4.1 Evaluation on CTCC Hefei SD. To evaluate the map matching performance, we compare MIFF with baseline models, shown in Table 2. HMM_MIFF improves 3.4% on PCMI and 58.9% on GE with respect to basic HMM, 2.6% on PCMI and 43.3% on GE when compared with HMM_RF. CTS is much specially designed for SD. In PCMI, CTS_MIFF is 6.5% and 5.7% higher than CTS and CTS_RF respectively. Similarly, on GE, CTS_MIFF decreased by 78.1% and 83.0%, respectively. It can be found that MIFF's fusion of the homogeneous traces cannot greatly improve the matching degree of the road segments, which could be represented by the improvement in PCMI. But MIFF can match the road segment that is not successfully matched closer to the actual road segment as much as possible. GE's improvement can prove this. In other words, the fusion of the cell tower level trace can reduce the deviation of the trace as a whole. When it comes to HMM_RF or CTS_RF, the fusion of the road level trace cannot effectively correct the result due to the limitation of the spatio-temporal uncertainty on map matching.

Table 2. Comparison of various models on CTCC Hefei SD

Models	PCMP	PCMR	PCMI	GE (m)
HMM	0.805	0.810	0.822	69.5
HMM_RF	0.820	0.837	0.828	58.6
HMM_MIFF	0.860	0.841	0.850	28.5
CTS	0.827	0.858	0.842	66.2
CTS_RF	0.860	0.830	0.849	69.5
CTS_MIFF	0.903	0.890	0.897	14.5

4.4.2 Evaluation on Various Transportation Modes. To evaluate the map matching performance on various transportation modes, we manually collect SD in Shanghai because it's hard to get ground truth from real-world SD. Four kinds of transportation modes are available in our Shanghai SD, which is shown in Table 1 with their data characteristics. Different data characteristics result in different model performance, thus the performance comparison on various transportation modes is necessary.

The experiments results are listed in Tabel 3. It shows that our model performs better in four transportation modes. Especially for train data, take HMM as an example, MIFF improves 47.34% on PCMI and 58.99% on GE. MIFF can also get a considerable improvement by taking CTS as a base model. Followed by bicycles, MIFF can also achieve significant improvement. On the contrary, there happens little improvement in the PCMI of MIFF with car data or bus data.

Next, we discuss some factors that have a significant impact on this evaluation.

- **Density of road network** The density of roads in the suburbs is relatively small, which leads to fewer intersections, so the human traces tend to be straight forward in one direction. This phenomenon also leads to an important problem, that is, once a certain road segment is matched incorrectly, it is difficult for the following road segment to return to the correct direction. Therefore, cell tower level fusion can correct the trace to a large extent, but in MIFF_RT, because part of the cell tower level trace is first matched to the wrong road segments, it is difficult to get the correct match result by road fusion. As for the urban area, the density of roads is very high, people often change roads to reach their destinations faster, which leads to little improvement in the PCMI of MIFF for car and bus.
- **Speed** With speeds increases, temporal frequency reduces, which may cause frequent path switch signals. Train data have low temporal frequency for the sparsity of the cellular towers nearby railways. Hence, the train data is high spatio-temporal uncertain for the high speed and low temporal frequency. Reduce spatio-temporal uncertainty by MIFF can greatly improve performance. so people tend to switch roads in a short time. When it comes to car and bus, fast-moving vehicles often switch roads in a short time. Therefore, a number of the roads that the vehicle traveled was not recorded by the cellular tower, which resulted in irreparable information losses. As we can see, MIFF cannot significantly improve the experimental results.

However, MIFF can achieve a greater improvement in bike than the other two baseline models. This is because the speed of bike is generally not fast and its trace can be captured actually.

Table 3. Comparison of various models on different transportation modes

Model		HMM	HMM_RF	HMM_MIFF	CTS	CTS_RF	CTS_MIFF
PCMI	Train	0.495	0.606	0.940	0.486	0.641	0.958
	Car	0.724	0.725	0.768	0.736	0.750	0.767
	Bus	0.543	0.561	0.592	0.580	0.596	0.601
	Bike	0.486	0.716	0.776	0.780	0.819	0.922
GE(m)	Train	69.5	58.6	28.5	66.2	69.5	17.5
	Car	78.9	68.4	65.7	74.1	74.4	60.4
	Bus	142.4	109.8	94.6	135.4	105.6	88.6
	Bike	173.7	81.2	61.3	42.7	33.6	10.4

4.4.3 Evaluation on Various Insights. According to the characteristics of SD data, we improve the homogeneous trace identification by two insights, looking for a regular pattern of the same users and finding co-movement between different users. These two ideas help us successfully find homogeneous traces in a large number of human mobility traces. But there are also differences between these two ideas, so we also designed experiments to compare the differences between them. Here we assume that the regular cell tower trace of the same user must occur three times or more, and the number of users in co-movement cell tower trace groups must greater than two. We divide traces from CTCC Hefei SD into two parts accordingly and use MIFF to evaluate the map matching performance, shown in Table 4.

It could be inferred that homogeneous traces composed of regular patterns of a single user that belong to the same group occur at different regular times, while a group of homogeneous traces formed by the co-movements between different users occurs in approximately the same period. Due to the uncertainty of cell tower, the user's cell phone may be connected to different cell towers when passing the same road segment at different times, and the temporal uncertainty will cause a great mismatch between SD traces belonging to the same homogeneous trace. Therefore, for the former, MIFF can reduce the probability of different cell towers belonging to the same road segment, so that GE can be reduced; while for the latter, the difference between different traces can be largely eliminated to obtain a correct match.

Table 4. Comparison of two insights in identifying homogeneous traces

Insight		HMM_MIFF	CTS_MIFF
Regular Pattern	PCMI	0.839	0.875
	GE (m)	20.66	12.41
Co-movement Pattern	PCMI	0.874	0.921
	GE (m)	33.8	25.6

4.4.4 Evaluation on Various Temporal Frequency. The performance gap on different transportation modes shown in Section 4.4.2 may be caused by different data characteristics, such as temporal frequency and speed. Here we conduct experiments to show the improvement of our model on various temporal frequencies, shown in Fig. 15. It shows that with the growth of time interval, our model improves too.

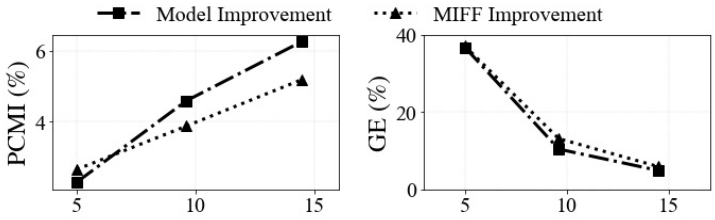


Fig. 15. Improvement for Various Time Interval

5 DISCUSSION

Lessons Learned: We summarize our lessons as follows.

- Our system can largely improve the performance of common map matching models like CTS and our Model without MIFF, as shown in Tabel. 2 by addressing the spatio-temporal uncertainty of SD.
- We found that generally, our model performs better on railway data than car data (Tabel 3). We believe it is brought by more co-moving people on the railway and the moving distances of railway users are longer than car users.
- We also found that our model performs better on car and bike modalities than bus (Tabel 3). It is because the routes of bus lines are usually winding, against the shortest path based model design.
- Our model performs better when time interval or spatial shifting factor increases, as shown in Fig 15.

These insights may provide some guidance for cellular network providers to match their detailed user demand distribution in the fine-grained deployment of 5G networks.

Generalization: The extensive results on signaling data and GPS data collected from two cities in China proved the generalizability of our method. Moreover, we argue that our work can also be generalized to other cities with similar data, such as cellular connection data, WiFi connection data, etc. We envision if the key factors such as statistic features of cellular data, and contextual information such as the distribution of cell tower/ access point and road networks are similar among cities, our results then can be generalized because they are the basis of all the above investigations.

Robustness: The CTCC Hefei SD dataset has SD of more than 3 million users within a month. SD trace may be generated by different transportation modes or even a mixture of different modes. In contrast, the generality of user behavior and large user coverage and spatio-temporal coverage is not available in other datasets with limited transportation modes, such as taxi dataset. But most SD traces cannot have the corresponding ground truth, which is also an unsolvable matter. With the aid of the additional DiDi dataset [5], we successfully obtained 13 groups of homogeneous traces to complete our experiment. These traces cover many different types of areas and different time periods within a day in the city and are produced by different modes of transportation. Experimental results demonstrate that MIFF performs well on comprehensive data sets. In Shanghai, our team manually collected SD data with ground truth through mobile devices. The amount of data collected manually in this part is relatively small. After data preprocessing, we get 70, 10, 7, 4 groups of homogeneous traces on the four transportation modes: train, car, bus, and bike. These trajectories generated by different users taking different modes of transportation have very different spatial and temporal distributions and cover different areas of cities and suburbs. The characteristics of different transportation modes are discussed in detail in Section 4.4.2. MIFF achieves great performance on different modes of data, which proves that our model is robust.

Privacy and Ethics: We have the opportunity to collaborate with the cellular service provider to analyze the cellular data for academic research purposes. As an agreement when signing cellular service contracts, all the users consent that their metadata will be used to perform analysis on anomaly detection, malicious cellular usage detection/blocking, access patterns, business opportunities, etc. In particular, location-specific cellular traffic demand modeling and prediction can enable many applications significantly related to cellular service quality

such as service coverage and load balancing. Furthermore, all users' IDs have been hashed into global identifiers, which cannot be leveraged to trace users. Most importantly, our ultimate objective is to improve cellular service quality by modeling fine-grained user locations, which is also consistent with users' willing. We envision the majority of users may not be against this work. As a result, our work is exempted from the institutional IRB process.

Data Sharing: Accessing empirical data sets is vital to mobility modeling and mobile system research, but such data sets are usually not available for fellow researchers due to the various real-world issues. As an initial step, we will release sample cellular data following the privacy protection schemes in [2] to enable further works upon our results. Details of the releasing data set are disclosed as below: i) time period: one week data from 2017-06-05 to 2017-06-11; ii) data amount: all data from a 100 km^2 region in downtown area; iii) data format: encoded user ID, the timing of the record, associated encoded tower ID, record type (specified as *sRequest*).

6 RELATED WORK

Similar Traces Search: The main work of measuring the similarity between traces is to calculate the distance. Dynamic Time Warping (DTW) [27] is the first algorithm applied to similar time sequences measurement. Subsequently, two threshold dependency algorithms, i.e., Edit Distance on Real sequence (EDR)[4] and Edit Distance with Real Penalty (ERP)[3], are proposed for capturing spatial semantics. While existing techniques are unable to cope well with temporal sparse traces. Most algorithms apply linear interpolation within two adjacent points, which may result in large errors in the geographic problem, e.g., Table 1. In contrast, we introduce a confidence concept as a weight while calculating the distance.

Map Matching: In recent years, several algorithms have been proposed to solve offline low frequency map matching with SD data. CTrack [19] fuses cellular fingerprints (cell tower observation) and additional sensor data collecting from mobile phones to perform map matching. However, it requires sensor data such as accelerometers, supervised data in the same place for pretraining, and more detailed cellular information such as neighboring towers and signal strength. SnapNet [15] proposes a thorough framework for SD data matching without additional sensors. While some common errors, such as U-turn and detour errors, are not taken into consideration. CTS [10] proposes an improved framework for SD data, containing noise filtering, state detection, and a HMM-based algorithm with U-turn and Direction changing penalties. While CTS does not consider the history and co-moving information, it uses additional sensor data, such as nearby 6 cell towers and their signal strength, cellular attributes, such as antenna direction and radiation angle of a sector. Additionally, many works are proposed to solve the low frequency map matching with GPS data. It is different from SD in two aspects: (1) the average geographic error of GPS is only 38.13m in our survey, which is much smaller than that of SD data, 262.27m; (2) the probability of connecting to the nearby cell tower is almost the same, which is greatly different from GPS. In GPS data, the observation is more likely beside the real point. Compared to the above, our system does not use any additional sensor data and cellular attributes, we utilize our framework to capture the history regular pattern and simultaneous information to improve any kinds of existing map matching for SD.

Spatio-temporal Data Analysis: Our work is also related to spatio-temporal data analysis. We use two metrics, i.e., the penetration rate and sparsity of data, to systematically organize all representative human mobility models as follows.

Table 5. Related work

	Low Penetration	High Penetration
Sparse Data	[26] [13]	[20] [8] [12]
Dense Data	[16] [31]	Our System MIFF

(1) For the low-penetration and sparse data quadrant, researchers use datasets such as survey data, data from bike sharing systems [26], and payment data to model and analyze human mobility. Thus the mobility models are not

representative at both crowd level due to low penetration rates and at the individual level due to sparse data, e.g., models driven by the bike data are only targeted at bike users at origin and destination only without detailed traces. (2) For the high-penetration and sparse data quadrant, data sources are usually with a high penetration rate covering almost all people while with relatively sparse data, such as CDR [12].

But even though almost every urban residents have mobile phones while these data can only be generated when they use their phones, e.g., making a phone call [20] [8]. As a result, these models can cover all residents but without detailed mobility traces. (3) For the low-penetration and dense data quadrant, such as GPS [31], they are one common data source used to analyze detailed human mobility since they have high updating frequency and high precision. However, they can only cover all residents participate in the systems, e.g., users with GPS devices in their cars, or residents taking taxis. (4) In contrast, our MIFF system is the first framework targeting high-penetration and dense data-driven human mobility modeling, which finds human history moving pattern, co-moving phenomenon, and proposes a more accurate map matching algorithm.

7 CONCLUSION AND FUTURE WORK

In this paper, we design a new framework MIFF to build human mobility extraction models based on unique features of Singaling Data. We conduct experiments on two kinds of SD, i.e., Hefei SD collected by the service provider and Shanghai SD collected by our group. To obtain more accurate matching results, we focus on reducing the negative impact of uncertainty by fusing homogenous traces.

The experiments considering various scenarios for different transportation modes, various insights, and time frequency. The results show that MIFF improves the map matching accuracy by about 5.0% on PCMI and 68.5% on GE on average compared to the state of the art map matching models without our framework. Besides, our framework can adapt to any new map matching algorithms in theory.

In the future, we will further study the co-moving phenomenon and regular patterns in public transportation. Besides, we will further study trajectory representation and road map embedding to improve the map matching algorithm under our proposed MIFF.

ACKNOWLEDGMENTS

This work is supported partly by the National Key R&D Program of China 2018YFB2100300, 2018YFB0803400, and National Natural Science Foundation of China (NSFC) 61925202, 61772046, and co-funded by DiDi GAIA Research Collaboration Program with trajectory data from DIDI.

REFERENCES

[1] Rajesh Krishna Balan, Khoa Xuan Nguyen, and Lingxiao Jiang. 2011. Real-time trip information service for a large taxi fleet. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*. 99–112.

[2] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data* 2 (2015), 150055.

[3] Lei Chen and Raymond Ng. 2004. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 792–803.

[4] Lei Chen, M Tamer Özsu, and Vincent Oria. 2005. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 491–502.

[5] Didi. [n. d.]. Didi dataset. https://www.didiglobal.com/.

[6] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban Population Modeling Based on Multiple Cellphone Networks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3 (2018).

[7] Ursula Gather and Verena Schultze. 1999. Robust estimation of scale of an exponential distribution. *Statistica Neerlandica* 53, 3 (1999), 327–341.

[8] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779.

[9] Bill Hillier, Alasdair Turner, Tao Yang, and H-T Park. 2009. Metric and topo-geometric properties of urban street networks: some convergences, divergences and new results. *Journal of Space Syntax Studies* (2009).
[10] Xingyu Huang, Yong Li, Yue Wang, Xinlei Chen, Yu Xiao, and Lin Zhang. 2018. CTS: A Cellular-based Trajectory Tracking System with GPS-level Accuracy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 140.
[11] Andreas Janecek, Danilo Valerio, Karin Anna Hummel, Fabio Ricciato, and Helmut Hlavacs. 2015. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE transactions on intelligent transportation systems* 16, 5 (2015), 2551–2572.
[12] Shan Jiang, Joseph Ferreira, and Marta C González. 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data* 3, 2 (2017), 208–219.
[13] Liang Liu, Anyang Hou, Assaf Biderman, Carlo Ratti, and Jun Chen. 2009. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference On*. IEEE, 1–6.
[14] Xiliang Liu, Kang Liu, Mingxiao Li, and Feng Lu. 2017. A ST-CRF map-matching method for low-frequency floating car data. *IEEE Transactions on Intelligent Transportation Systems* 18, 5 (2017), 1241–1254.
[15] Reham Mohamed, Heba Aly, and Moustafa Youssef. 2017. Accurate real-time map matching for challenging environments. *IEEE Transactions on Intelligent Transportation Systems* 18, 4 (2017), 847–857.
[16] NYC Department of Transportation. [n. d.]. New York City Camera. http://dotsignals.org. 2017.
[17] Zhou Qin, Zhihan Fang, Yunhuai Liu, Chang Tan, Wei Chang, and Desheng Zhang. 2018. EXIMIUS: A measurement framework for explicit and implicit urban traffic sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 1–14.
[18] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2014. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 5–14.
[19] Arvind Thiagarajan, Lenin Ravindranath, Hari Balakrishnan, Samuel Madden, and Lewis Girod. 2011. Accurate, low-energy trajectory mapping for mobile devices. (2011).
[20] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. 2009. VTrack: energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM conference on embedded networked sensor systems*. ACM, 85–98.
[21] Guanfeng Wang and Roger Zimmermann. 2014. Eddy: an error-bounded delay-bounded real-time map matching algorithm using HMM and online Viterbi decoder. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 33–42.
[22] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. 2012. Quantifying the impact of human mobility on malaria. *Science* 338, 6104 (2012), 267–270.
[23] Wikipedia. [n. d.]. 2014 Shanghai stampede. https://en.wikipedia.org/wiki/2014Shanghaistampede/. Dec 31st, 2014.
[24] Wikipedia. [n. d.]. C114 China Communication Network. http://www.c114.com.cn/market/220/a1042288.html. Freb 1st, 2018.
[25] Hao Wu, Ziyang Chen, Weiwei Sun, Baihua Zheng, and Wei Wang. 2017. Modeling Trajectories with Recurrent Neural Networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3083–3090. https://doi.org/10.24963/ijcai.2017/430
[26] Zidong Yang, Ji Hu, Yuanchao Shu, Peng Cheng, Jiming Chen, and Thomas Moscibroda. 2016. Mobility modeling and prediction in bike-sharing systems. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 165–178.
[27] Byoung-Kee Yi, HV Jagadish, and Christos Faloutsos. 1998. Efficient retrieval of similar time sequences under time warping. In *Data Engineering, 1998. Proceedings., 14th International Conference on*. IEEE, 201–208.
[28] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. 186–194.
[29] Kai Zhao, Jie Feng, Zhao Xu, Tong Xia, Lin Chen, Funing Sun, Diansheng Guo, Depeng Jin, and Yong Li. 2019. DeepMM: Deep Learning Based Map Matching with Data Augmentation. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19)*. Association for Computing Machinery, New York, NY, USA, 452–455. https://doi.org/10.1145/3347146.3359090
[30] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. 2010. Understanding Transportation Modes Based on GPS Data for Web Applications. *ACM Trans. Web* 4, 1 (2010).
[31] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 312–321.
[32] Fangzhou Zhu, Chen Luo, Mingxuan Yuan, Yijian Zhu, Zhengqing Zhang, Tao Gu, Ke Deng, Weixiong Rao, and Jia Zeng. 2016. City-Scale Localization with Telco Big Data. 439–448. https://doi.org/10.1145/2983323.2983345