CellSense: Human Mobility Recovery via Cellular Network Data Enhancement

ZHIHAN FANG, Rutgers University, USA
YU YANG, Lehigh University, USA
GUANG YANG, Rutgers University, USA
YIKUAN XIAN, Rutgers University, USA
FAN ZHANG, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd.
DESHENG ZHANG, Rutgers University, USA

Data from the cellular network have been proved as one of the most promising way to understand large-scale human mobility for various ubiquitous computing applications due to the high penetration of cellphones and low collection cost. Existing mobility models driven by cellular network data suffer from sparse spatial-temporal observations because user locations are recorded with cellphone activities, e.g., calls, text, or internet access. In this paper, we design a human mobility recovery system called `CellSense` to take the sparse cellular billing data (CBR) as input and outputs dense continuous records to recover the sensing gap when using cellular networks as sensing systems to sense the human mobility. There is limited work on this kind of recovery systems at large scale because even though it is straightforward to design a recovery system based on regression models, it is very challenging to evaluate these models at large scale due to the lack of the ground truth data. In this paper, we explore a new opportunity based on the upgrade of cellular infrastructures to obtain cellular network signaling data as the ground truth data, which log the interaction between cellphones and cellular towers at signal levels (e.g., attaching, detaching, paging) even without billable activities. Based on the signaling data, we design a system `CellSense` for human mobility recovery by integrating collective mobility patterns with individual mobility modeling, which achieves the 35.3% improvement over the state-of-the-art models. The key application of our recovery model is to take regular sparse CBR data that a researcher already has, and to recover the missing data due to sensing gaps of CBR data to produce a dense cellular data for them to train a machine learning model for their use cases, e.g., next location prediction.

CCS Concepts: • **Networks** → *Location based services*; • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Cellular Network, Human Mobility, Signaling

ACM Reference Format:
Zhihan Fang, Yu Yang, Guang Yang, Yikuan Xian, Fan Zhang, and Desheng Zhang. 2021. CellSense: Human Mobility Recovery via Cellular Network Data Enhancement. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 100 (September 2021), 22 pages. https://doi.org/10.1145/3478087

Authors' addresses: Zhihan Fang, Rutgers University, Piscataway, NJ, 08854, USA, zhihan.fang@rutgers.edu; Yu Yang, Lehigh University, USA; Guang Yang, Rutgers University, USA; Yikuan Xian, Rutgers University, USA; Fan Zhang, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd. Desheng Zhang, Rutgers University, USA, desheng.zhang@cs.rutgers.edu.

2474-9567/2021/9-ART100 $15.00
https://doi.org/10.1145/3478087

1 INTRODUCTION

Sensing human mobility in term of fine-grained locations is of great importance for ubiquitous computing [4] [66], location-based services [65], urban planning[54] [55], and recent pandemic mitigation [52]. Most existing sensing systems to capture human mobility suffer from either high deployment cost [34] or low spatial coverage [33]. Recently, thanks to the ubiquitous cellular devices, the study of human mobility sensing on cellular networks has gained significant attention because of the high penetration rate of cellphones and the low marginal cost to collect cellphone data.

The cellular service providers collect Cellular Billing Records (CBRs) for billing purposes such as phone calls, messages, and internet access, which can be re-purposed to sense users' locations and benefit the cellular providers for their use cases, e.g., cellular network capacity modeling. Compared with Call Detail Records (CDR) data, which mostly refers to records for users' phone calls and message calls, CBR is a more general dataset covering both call records and internet connection records. Relying on various CBR datasets collected by network operators, human mobility sensing is extensively investigated by fellow researchers [28] [41] [64]. Existing works mainly focus on either (1) collective mobility such as flows [17] and travel time [61], or (2) individual mobility such as commuting patterns [27] and user spatial profiling [28] based on statistical methods, e.g., Hidden Markov Model, or Conditional Random Field, which leads to satisfactory performance [65]. We summarize the existing work on cellphone billing records (CBR) in table 1.

Table 1. Related Work Using Cellphone Billing Records (CBR)

Name	Location	# Days	Volume
[28]	NY & LA	91	585K users
[25]	NY & LA	78	168K users
[26]	NY & LA	140	352K users
[37]	New York	91	250K users
[3]	Ivory Coast	150	50K users
[35]	China	—	100K users
[8]	Shanghai	14	642K users
[30]	Singapore	14	3.17M users
[67]	Shenyang & Dalian	48	3M users
[22]	Dhaka City	30	2.87M users
[51]	Paris	21	800M records
[17]	Shenzhen	—	10.2M records
[23]	Milan	60	319M records
[11]	Italy	67	17K trajectories

Recent advance of Machine Learning, especially deep learning, has the potential to further improve human mobility modeling accuracy, e.g., prediction accuracy. However, the challenges of using cellular network data for machine learning models is that users' CBR are often too sparse on temporal dimension. This is because these CBR data are mainly collected for billing purposes, e.g., phone call, text, and internet access, but most of people have long period time of day without any CBR data because of no activities. For example, in our CBR dataset, the average time with no activities is around 12 hours, which occupy 75% of people's everyday time if considering 8-hour sleeping. As a result, it is challenging for researchers to explore the advance of machine learning to improve human mobility models because of CBR data sparsity. It motivates us to design a recovery model to infer historical missing observations in the gaps between raw CBR data of individual users. With such a recovery model, the fellow researchers (e.g., the ones in table 1) can enhance their legacy sparse CBR data to obtain dense CBR data to train predictive machine learning models for their use cases.

Existing works address the above sparsity challenge from three aspects. (i) Some work integrates CBR data with other mobility data sources, e.g., transportation data, with methods such as cotraining [17] or multi-view learning [64]. Studies in this category require extra data sources for calibration, which are not always accessible by other researchers with CBR data. (ii) Other similar works include map matching from cellular towers such as ctrack [50], coSense [57], which achieves continuous location sensing from connection with cellular towers; location positioning by combining cellular signal strength distribution from nearby towers and CDR data [18]; However, these works require dedicated data collection such as continuous active cellphone connections or cellphone interior sensor data. (iii) Admittedly, in theory, some work can infer the missing observations to fill sensing gaps in cellular data by classical regression models. Those work are either built on theoretical models with statistical assumptions [42], or lack details of implementation and evaluations [16]. However, little work has been done in this category because it is challenging to obtain large-scale ground truth data for model validation.

Different from existing works, data recovery on CBR introduces two unique challenges: (i) uncertainty in individual mobility caused by insufficient sensing data and irregular cellular usage patterns. This is because the CBR is passively collected for billing purposes instead of active mobility tracking, and therefore the data quality highly depends on cellular usage patterns of individual users. (ii) in practice, it is almost impossible to calibrate the uncertainty with external data sources such as transportation data because users have different anonymous IDs in different datasets due to encryption, and fine-grained mobility is hard to be matched with data fusion from different data sources at the individual level. To address the two specific challenges in CBR data recovery, our recovery methodology is built with two key techniques, i.e., collective individual mobility calibration in a single data source, and stage-based bidirectional learning. We extract collective mobility features (e.g., travel time, based on CBR data from all users) and apply the collective mobility features as calibration factors in individual mobility inference to address the uncertainty in individual CBR data. For individual users, instead of focusing on single records, we divide users into a staged distribution of observed and unobserved stages. Based on continuous records in the observed stages, we infer user status such as transportation modalities, which are used in the learning phase.

To rigorously evaluate our recovery model, we explore a new opportunity based on the upgrade of cellular infrastructures to obtain the ground truth data. In particular, recently, the cellular service providers have been logging signaling data, in addition to the CBR. Different from the CBR for billing purposes, the signaling data log the interactions between cellphones and cellular towers, e.g., attaching, detaching, paging, etc, even without billable activities. These signaling data were not typically logged and stored by the operators before due to the large data storage without immediate use cases. But given the development 5G and AI use cases, the service providers have realized the value of such data and started to invest to collect these signaling data. We have been working with one of cellular operators in Hefei City, China to access these signaling data as the ground truth for a recovery model we build. The key application of our recovery model is to take legacy sparse CBR data that a researcher has (e.g., the ones in in table 1), and to recover the missing data due to sensing gaps of CBR to produce a dense cellular data for them to train a machine learning model for their use cases, e.g., next location predictions.

The problem we are focusing is straightforward: *Can we recover missing records in CBR to improve data quality without access to other data sources?* To address the above problem, we design and implement a human mobility recovery system named `CellSense` with two key components. an individual-independent component for collective mobility modeling, and an individual-dependent component for context-aware individual mobility modeling. We summarize our contributions as follows.

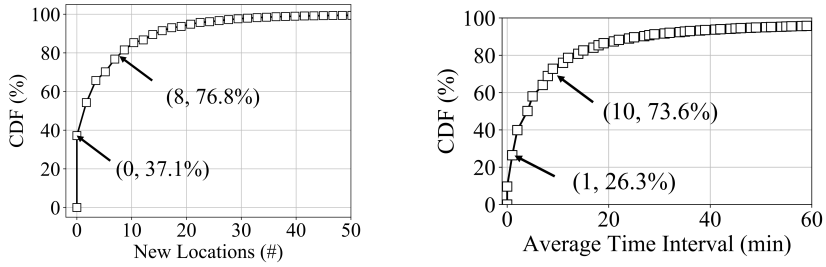
- To our knowledge, we conduct the first study `CellSense` to infer sensing gaps in cellular billing records (CBR) of cellular networks, which benefits both fine-grained human mobility and existing researchers with sparse CBR. Our study is based on CBR (i.e., for phonecall, message and internet access) of a city-scale cellular network in Hefei city, China covering around 3.37 million active users (around 40% penetration rate

in the city). Under the consent of our collaborators, we will share one week of sample data including both cellular billing records and signaling records so fellow researchers can validate and build upon our work.
- To address the prediction uncertainty in individual mobility, we have three key system design and technical contribution in `CellSense`: (i) we integrate collective and individual mobility in one uniform learning framework to solve uncertainty of individual-level human mobility; (ii) We normalize and embed the contextual information from heterogeneous data sources in the learning framework to calibrate the mobility learning; (iii) Instead of focusing on single directional learning (forward direction) in most existing studies on human mobility, we consider mobility from two directions, e.g., past and future mobility observations, and successfully apply bidirectional learning in `CellSense`.
- We evaluate `CellSense` with a separate data source, e.g., signaling data, as the ground truth, which are collected internally by cellular network and different than CBR, i.e., the input of `CellSense`. The evaluation results show that `CellSense` achieves 35.3% improvement on performance compared with state-of-the-art methods.

2 MOTIVATION

2.1 Challenges

2.1.1 New Locations. Even though human mobility shows regular patterns such as commuting between home and work locations, there exists uncertainty in human mobility. As shown in Fig. 1a, we profile users' visited locations with two weeks of CBR and then study the irregularity of users by new locations in the following week. we found only 37.1% users move among existing locations and around 23.2% users visited more than 8 new locations. The new locations introduce challenges for human mobility recovery due to the lack of historical observations.



(a) New Visited Locations (b) Average Interval

Fig. 1. Challenges of CBR Data for Mobility Recovery

2.1.2 Sensing Granularity. Different from sensors such as GPS devices, which passively sense human locations with constant time intervals, e.g., 5 seconds, cellphone devices rely on active user activities for location sensing. As a result, the temporal sensing granularity of cellular billing records is non-uniform. The average sensing granularity, i.e., time interval between two CBR records, differs among users, which is caused by different cellphone usage patterns of users. As shown in Fig. 1b, the average time interval between two CBR records is less than 1 minute among 26.3% users while more than 10 minutes in top 26.4% (100%-73.6%) users. The non-uniform and irregular sensing granularity introduces new challenges for human mobility recovery.

2.2 Opportunities

2.2.1 Increasing Demand for Cellular Services. With the increasing demand of high-quality, ubiquitous internet access, cellular networks become one of the most promising ways to sensing human mobility in the city since mobile devices such cellphones and tablets rely on cellular networks for internet services. According to Statista, cellphone users have been increasing from 4.3 billion in 2016 to 4.8 billion in 2020, and in addition smartphone

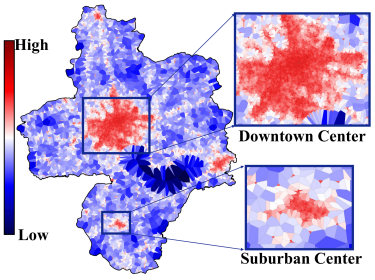


Fig. 2. User Density Distribution

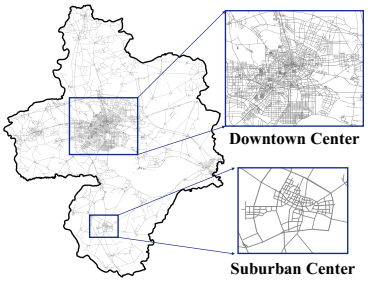


Fig. 3. Road Networks

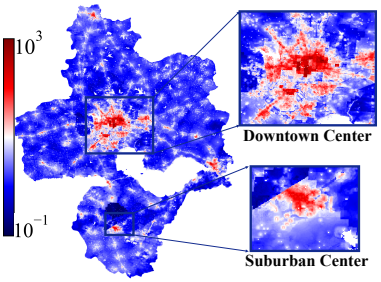


Fig. 4. Population

users increases from 2.5 billion to 3.5 billion [38]. More importantly, due to the increasing demand of internet services, users spend more time on cellular services for internet access. Compared with 62 minutes of average daily usage time and 0.38 GB of average handset data traffic per user per month in 2015, users spend 358.2 minutes per day on their cellphones and use 7.82 GB cellular networks per month for data transmission in 2019 [1]. In particular, most users use cellular networks for data transmission in outdoor settings when they have no access to indoor WiFi. We divide user behaviors into two stages, i.e., an observed stage with continuous user location observations and an unobserved stage without any cellular billing records. The increasing demand of cellular services generates enough CBR data to (i) infer individual travel features such as speeds and transportation modes in the observed stages, and (ii) provide high spatio-temporal coverage for collective travel features such as travel time on roads.

2.2.2 Contextual Information. To provide better service experience, in recent years, cellular providers start to collect users' contextual information with incentives, e.g., data package rate discounts. In return, users consent to provide high-level contextual information for social good, backend analysis and metric monitoring, e.g., subscription plan information, active time on certain apps, etc. The contextual information includes a wide range of demographic factors describing users such as user age, car owner or not, which are used for the user profiling for the cellular provides' new services and products. Such demographic information is correlated with the usage and mobility patterns of users. Therefore, in this paper, we collaborate with cellular providers and integrate such contexts in our mobility modeling. The details of our data are given in the following section.

3 DATASETS

Cellular Billing Records (CBR): We are collaborating with one of three major cellular network operators in China and have offline data access to its CBR data under NDA (Non-disclosure agreement). Please see our discussion section our privacy protection, consent and ethics consideration. This cellular network has 23,704 cellular towers providing services for 3.37 million users in a Chinese city Hefei, the capital of the Anhui province. It generates 20GB daily usage records, which we use for our mobility recovery. CBR contain the location and time information when a user uses cellular services including phone calls, messages, and internet services. Each CBR record has 5 attributes. A sample record and some statistical information is given in Table 2.

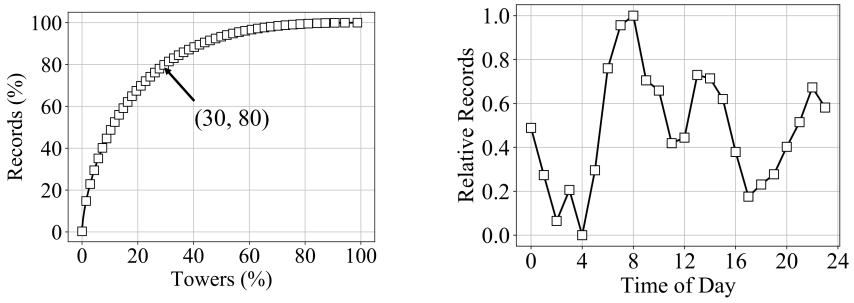
The details of each attribute are given as follows.

- *Timestamp*: Date and time when the record was generated. The precision of the time is millisecond.
- *User ID*: A unique encrypted identification of the cellphone user who generates the record.
- *Tower ID and Tower Location*: A unique identification of a cellular tower and its corresponding GPS location with a longitude and latitude.
- *Connection Type*: The types of connections, i.e., phone calls, messages, or internet services.

Table 2. Cellular Billing Records

Field	Value
Timestamp	2017/06/12 00:23:55.357
Tower ID	110613042
Tower Location	116.9855024, 31.7906872
User ID	3B8xZpNZJpTjfOnwYbcdyA==
Connection Type	4G
Number of Daily Records	267 million
Number of Daily Users	3.37 million
Number of Cell Towers	23,704

We study the spatial and temporal distribution of CBR for a preliminary analysis. On the spatial dimension, a Voronoi partition is applied to estimate spatial coverage of cellular towers [17]. As shown in Fig. 2, we found a higher user density in downtown compared with surrounding suburban areas. The quantitative results are reported and compared in Fig. 5a, where we found a unbalanced distribution of records on towers. 80% records are concentrated on 30% cellular towers, which are mostly located in the downtown areas of the city. This is because there is a higher user density and business activities in the downtown areas compared with suburban areas. It indicates an uneven distribution of user locations for mobility inference. We further study the temporal distribution of cellular activities in Fig. 5b. We found three peak usages at 8:00, 13:00, and 22:00, which are corresponding to the morning peak hour, lunch time, and evening peak hour of users.



(a) Records (b) Users

Fig. 5. (a) Spatial Distribution of CBR on cellular towers (b) Temporal Distribution of CBR during One Day

User Contextual Information: We are also granted the access to the user contextual data collected from subscription data and cellular operator metadata. The subscription data come from a new commercial pattern emerging in China, where IT companies cooperate with cellular operators to dedicate cellular data plans for specific apps or services with a lower cellular data price, e.g., providing a 5GB exclusive data package which can only be used for certain car insurance apps or video apps. Then these contextual data can indirectly provide us with more behavioral information about users. For example, "car" tags can indicate this user may own a car, which further indicates that this user may use a navigation app or listen to streaming music, resulting in more cellular data consumption. In total, all subscribed cellular users in our CBR data have tags. Around 80% of users have fewer than 5 tags, the maximum number of tags for a user is 37, and the average number of tags for all users is 4. In summary, the six categories and statistics are shown in Table 3. The first column shows the categories; the second column denotes the percentage of corresponding tags of this category out of all the tags; the last column presents the percentage of users with tags of this category out of all the users. For clarification and

brevity, here we show some representative examples of tags for categories. For the phone usage category, typical tags include video, music, etc; for the car-related category, typical tags include car-related properties such as the user has car(s); for the demographic category, typical tags include gender, age range (e.g., the twenties), etc; for the services category, typical tags are related to services in life, such as Internet services. These tags from cellular service providers were collected with incentives, e.g., rate discount, to improve user experience and network efficiency with better understanding of correlation between usage patterns and their contextual information. We gives some detailed distributions on major tags as follows.

Table 3. Categories and Statistics

Categories	Tag/Tags(%)	Users/Users(%)
Phone usage	38.64	87.25
Car related	35.87	83.68
Demographic	17.91	46.74
Services	3.95	12.19
Average number of tags per user: 4		
Number of users with tags: 3.37 millions		

Road Network: Road networks describe the topological structure of human mobility in cities and highly correlated with spatial traces. The road networks of Hefei city are collected from OpenStreetMap [21], which include 4,9603 road segments and have a length of 12,813 km. The road network distribution in Fig. 3 complies with the user density in Fig. 2, i.e., we found more users in places with a denser road segments.

Population: Another impact indicators for human mobility modeling is the static population distribution based on census data [17]. In general, it is more possible for a user to visit places with a higher population distribution. We visualize population distribution in Fig. 4 based on census data collected by the Worldpop project [19]. Similar to the road network distribution, we found a high correlation between user density distribution and population distribution when comparing Fig. 2 with Fig. 4.

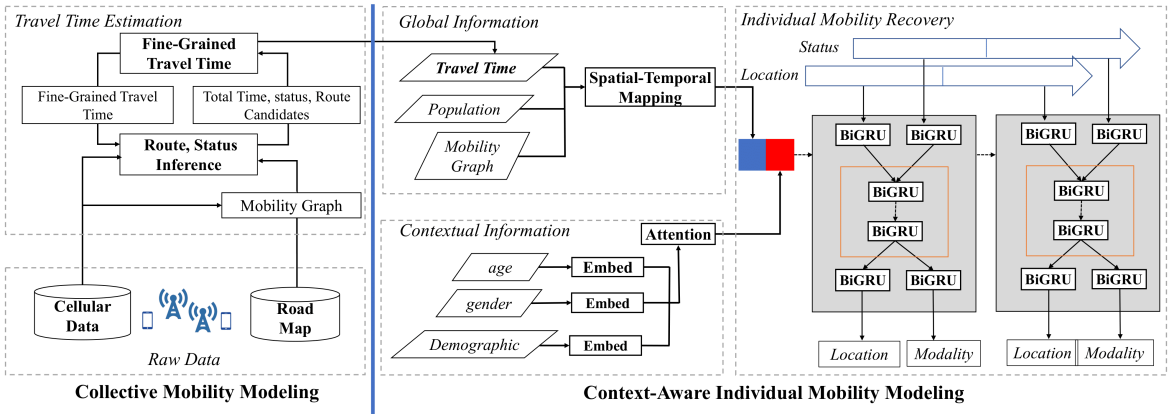


Fig. 6. Framework

4 METHODOLOGY

In this section, we elaborate on design of `CellSense`. We introduce the system overview, followed design details of the system.

4.1 System Overview

In our design, we utilize the two opportunities to address the two challenges described in the motivation section with two key components as shown in Fig. 6: (i) an individual-independent component for collective mobility modeling; (ii) an individual-dependent component for context-aware individual mobility modeling.

4.2 Preliminary

4.2.1 Spatial and Temporal Granularity. We use Voronoi partition to estimate the coverage of cellular towers [17] since users are connected to the nearest towers. On the temporal dimension, we divide time into 5-minute time slots. If a user is connected to more than one tower during the same time slot, we set the location in the time slot as the most frequent tower. A user mobility is described by a sequence of towers in 5-minute time slots.

4.2.2 Mobility Graph Construction. In the first step, since road networks describe the topology of user mobility but our sensing granularity is on cellular towers, we construct a *mobility graph* by combining both road networks and tower locations. Specifically, a *mobility graph* is defined as $G = (V, E)$ where V is tower locations and E is the collection of edges connecting cellular towers. As shown in Fig. 7, when a road connects two adjacent covered areas of towers, e.g., tower 1 and tower 2, an edge will be added into E.

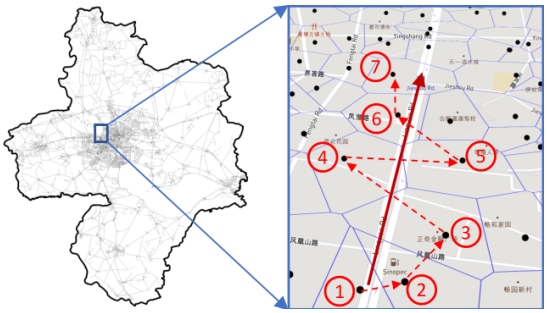


Fig. 7. Mobility Graph Construction

4.2.3 Ping-pong Effects and Outliers Outliers and noise exist in user records, which are caused by many factors such as load balancing [44] [15] and Ping-Pong effects [24] [41]. Even those outlier records are always ignored and invisible for collective mobility analysis, they are identifiable on individual mobility when we conduct a correlation analysis on a user's current locations with his previous and later status. We give an example of such outliers in Fig. 8 where the red points are locations of towers. A user drives on highway of the city and follows the mobility graph structure. However, the user is connected to a tower which is far from his actual trace due to ping-pong effect. We identify outlier records by combining mobility graph with heuristic features such as travel speed. During the trip, for a user u, we sort his records r with time and then calculate travel speed $s_{i\rightarrow(i+1)}$ of every adjacent pair of records r_i and r_{i+1}. For all travel speed $s_{i\rightarrow(i+1)}, \forall\ i$ during a time period, e.g., 1 hour, we calculate their mean and variance. We calculate z-score of speed by $z = \frac{|s_{i\rightarrow(i+1)} - \bar{s}|}{\sigma}$, and remove r_{i+1} if the z-score $s_{i\rightarrow(i+1)}$ is less than a predefined threshold with a certain confidence interval.

4.2.4 Stay Point Detection and Trip Segmentation. Stay points are locations where a user stays for a certain time. Those locations are always important PoI (point of interests) for spatial profiling, such as home and work locations [28]. More importantly, stay points are segmentation boundaries of logistic trips. We identify stay points based on travel speed and time, e.g., travel speed is 0 during a certain time period. Based on the stay points, we

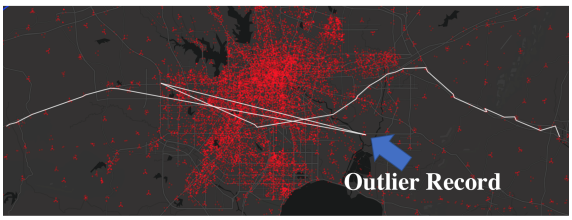


Fig. 8. Outlier Records

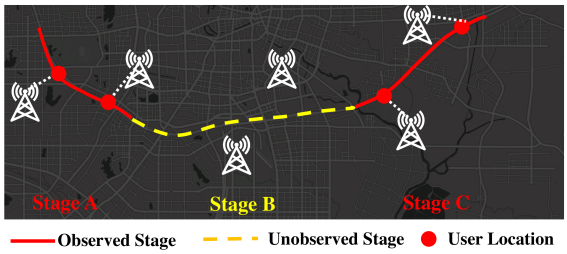


Fig. 9. Staggered Stages

segment user records into logistic travel trips. If both origin and destination of a trip are the same stay point, we label all user missing locations during this trip as the stay point.

4.2.5 Observed and Unobserved Stages. Based on analysis, we found cellular users use cellular service in a short continuous time period during a trip and then disconnect for a while. For instance, a user will check emails or send instant messages with friends when riding on bus. The users' locations are observed in cellular networks during this time period. When no cellular connection is established with nearby towers, users' locations become invisible. Therefore, we divide a user trace into two stages, i.e., an observed stage with cellular activity and non-observed stage with no cellular activity. As a result, a user's trip consists of staggered distribution of observed and unobserved stages as shown in Fig. 9. Our target is to infer locations of users at certain time in unobserved stages. In the observed stage, we can infer transportation modality of cellular users by thresholds of speed since we have continuous observations. Specifically, we label user modality in the observed stage into three modalities including walking, bikes and vehicles based on the travel speed. We initialize modality of users in unobserved stage from its last observed stage and then dynamic adjust their modality with model iteration.

4.3 Collective Mobility Modeling

4.3.1 Target. To infer missing records in unobserved stages, i.e., locations of users at certain time, it is essential to estimate both travel time and routes of users during the unobserved stages. However, the individual users have limited number of observed records. Many edges on the mobility graph lack direct observations from users for travel time estimation. Therefore, it is challenging to directly estimate travel time and routes for individuals. We first estimate the average travel time with different modalities and then integrate the average travel time with individual features, which is introduced in the individual-dependent mobility modeling. Instead of focusing on individual mobility modeling, we estimate user average travel time on the mobility graph under different transportation modalities.

4.3.2 Design. For edges with direct observations, the travel time can be directly estimated by the statistic mean. However, for edges with sparse observations, travel time estimation is challenging due to the lack of direct observations even users travel on those edges in their unobserved stages. To solve both cases, it is required to infer travel time from both direct observations (observed stages) and indirect observations (unobserved stages). In general, total travel time on any trip can be expressed by the following formula where $X^T = (x_1, x_2, \ldots, x_i, \ldots, x_n)$ and x_i is the travel time on edge e_i and $P = (p_1, p_2, \ldots, p_i, \ldots, p_n)$ is the route indicator where $p_i = 1$ when user pass edge e_i and otherwise $p_i = 0$.

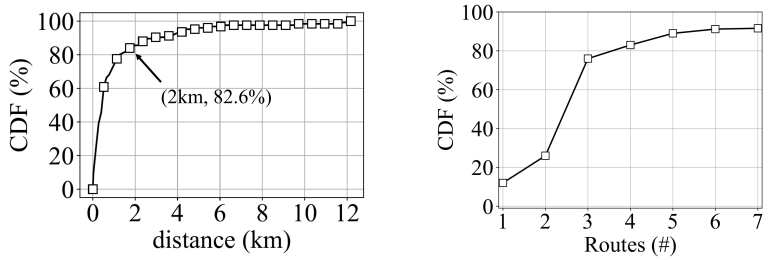
$$\tau = X^T P \tag{1}$$

For observed stages, the travel time on edges can be directly inferred from user observations. For unobserved stages, only total travel time is known, which can be inferred by the time difference of last and later observations.

We take two steps to estimate the travel time from unobserved stages: (i) route inference; and (ii) edge travel time inference as follows.

4.3.3 Route Inference. We infer detailed moving route of users on mobility graph based on distance and travel time constraints. Theoretically, there are an infinite number of routes between two locations on the mobility graph. To avoid the extra computing cost and make the problem feasible, we only focus on non-returning travel routes, in which a user will not travel the same edge twice. We validate this assumption from users' observed stages and found around 98.3% trips covers no-duplicate edges in one trip. We apply the Depth-First Search (DFS) algorithm to search for possible route candidates between two observed locations before and after an unobserved stage. We set a visited indicator for each search and prune all returning routes. As shown in Fig. 10a, we found travel distance of 82.6% unobserved stages is less than 2 kilometers.

To further reduce the number of route candidates and identify users' actual travel routes, we apply several heuristic factors to prune route candidates: (i) travel speed should be in a reasonable range, e.g., less than 80km per hour in downtown area and 150km per hour in suburban area; (ii) because passengers normally choose routes with short distance or travel time, we apply another heuristic constraint: travel distance should be less than a certain threshold, e.g., twice of the shortest distance. With both pruning methods, the number of route candidates is less than 3 for 79.3% unobserved stages as shown in Fig. 10b.



(a) Distance Interval

(b) Route Candidates

Fig. 10. Distance and Routes in Unobserved Stages

4.3.4 Edge Travel Time Inference. In the unobserved stages, for Equation 1, the total travel time τ and travel route P are known from route candidate searching and pruning. Instead, edge travel time X remains unknown. To estimate the travel time X, we present the edge travel time X with hidden parameters θ, i.e., means and variances for Gaussian distributions. For one edge and one modality, the set of parameters includes mean and variance to present a Gaussian distribution for travel time. With the presentation, our target is to estimate a set of parameters θ, which maximize likelihood estimation of τ in Equation 1. To solve the problem, we apply an EM (Expectation-Maximization) algorithm [63] [62]. First of all, we initialize the set of parameters from surrounding edges that have direct observations. Second, we update edge travel time with an iterative EM algorithm with two steps. In the E-step, we generate samples from each edge on the route of CBR according to the current edge travel time distribution. Each sample contains two elements, the travel time x_i and the corresponding probability in the distribution. In the M step, we update the parameters of edge e with new samples.

4.4 Individual Mobility Modeling

4.4.1 Global Information. The global information module fuses the estimated travel time from collective mobility modeling, population and mobility graph. Human mobility is highly correlated with population distribution in cities, e.g., higher travel demand in regions with high population density [59]. Even though population distribution can be modeled from cellular network data, previous work has revealed that user distribution from single networks could be biased to reflect the real accurate population distribution [15]. Therefore, we use a

separate census-based data source for population distribution [19] . The mobility graph and travel time are created and estimated from collective mobility modeling. The population distribution is directly inferred from the worldpop dataset. Since population and mobility graph are static information, which do not change during a short time period, we apply a spatial mapping on this two datasets, e.g., dividing cities into small Voronoi regions and then using the mobility graph and population density in the same regions where individual users are located. Travel time dynamically changes with time and locations, so both spatial and temporal mapping is applied. Specifically, we calculate the average travel time in Voronoi regions during different time slots of a day. When modeling individual mobility, we aligned the travel time in the same region and same time slot with individual users. Then the output of spatial-temporal mapping is the concatenation of global information aligned in regions and time slots.

4.4.2 Contextual Information Embedding. The contextual information includes the demographic factors of users. We found those information has an impact on individual mobility patterns. The impact of factors differs among users. For example, it is more possible that a car owner will drive a car for the commuting purpose but the possibility among users are different. If the user lives in suburban areas, the car usage will be more frequent for the commuting purpose compared with users living in downtown areas with a denser coverage of public transportation. To capture those differences of hidden correlation, we adopt an attention mechanism to automatically learn the weights of factors because attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of dependencies [53].

4.4.3 Bidirectional Mobility Learning. We apply a Bi-directional GRU model to integrate users' mobility inference with collective mobility patterns with personal information. Specifically, we integrate the mobility graph, travel time on graph, and personal contextual information with location estimation. Due to the spatial-temporal nature, the individual mobility presents a high correlation with the spatial and temporal information.

Recurrent Neural Network (RNN) is especially suitable to capture the temporal and spatial evolution of human moving. Compared with a regression model, which restricts a constant relation between input and output, e.g., a polynomial relation, RNN presents higher flexibility on hidden relations. Besides, the configuration flexibility makes it suitable to integrate spatial and temporal dependency. However, previous studies [49] have shown that traditional RNNs fail to capture the long temporal dependency for the input sequence due to the vanishing gradient and exploding gradient problems. To address these drawbacks, gated recurrent units (GRU) are a special RNN architecture for sequence labeling and prediction tasks [10]. Compared with Long-Short Term Memory (LSTM) unit, GRU is simpler to compute and converges faster. Therefore, we apply a time series learning GRU model combined with collective mobility, contextual information and user historical records to capture individual mobility dynamics. Specifically, the input of the GRU model is the concatenation of collective mobility, contextual information, and individual historical records. For static information such as age, gender, we directly copy the same values in different positions of inputs for a same user. Fig. 11 illustrates the internal structure of GRU cell, which consists of two gates, i.e., reset gate r and update gate z, and h and $\tilde{h}$ are the activation and the candidate activation. We can use the following formula to present the learning process. In the formula, x_t is the input vector at time t, h_t is the out vector, $\tilde{h}_t$ is the candidate activation vector, z_t is the update gate vector, r_t is the reset gate vector. W, U and b are parameters to learn in the training process. σ_g is a sigmoid function as the

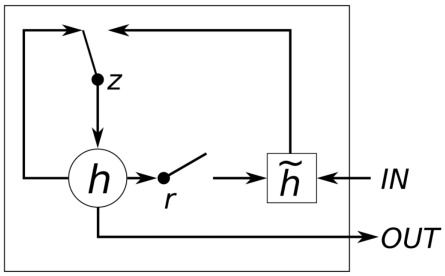


Fig. 11. GRU Cell

activation function which functions as a gate, i.e., 1 with large values and 0 with small values.

$$
\begin{aligned}
z_t &= \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \\
r_t &= \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \\
\hat{h}_t &= \Phi_h(W_h x_t + U_h(r_t \odot h_{t-1} + b_h) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t
\end{aligned}
\tag{2}
$$

A single direction GRU is widely used in prediction tasks, in which the input vectors only contain historical information. Different from prediction tasks, we found users' current locations are highly correlated with both his previous and later locations. Because our target is to infer sensing gaps instead of focusing on prediction, we apply a bi-directional GRU to utilize users' past and future mobility records for inference. To infer a sensing gap, a two-direction information will be learned and stored in hidden states. In the forward pass, the past mobility patterns will be learned and encoded in hidden states. In the backward pass, the future mobility patterns will be reversed. The reversed mobility patterns will be learned and encoded in hidden states.

We initialize the transportation modality of users in the unobserved stages from last observed stages. Users' transportation modality is highly correlated with users' moving speed, routes and locations. In general, users keep the same modality in unobserved stages from last observed stages because distance and time of unobserved stages are small and not enough for modality changes as shown in Fig. 10a. As shown in Fig. 9, the modality of stage B (unobserved stage) will be initialized with modality of stage A (observed stage with direct continuous observations). To utilize the prior information and dynamically update user modality in the training process, we adopt a multi-task learning mechanism in the training process. Specially, we apply an encoder layer to embed personal mobility records and transportation modality; a decoder layer to decode the embedded locations and transportation modality. We apply two shared hidden layers between the encoder and decoder. At each iteration, the updated modality will be fed into training and prediction of next time slots for calibration.

4.4.4 Model details. In our GRU model, after careful tuning, the number of time slots used as the input is 9 and the output at a time is 1 location. To minimize the training cost, we use the euclidean distance as the loss function because it is almost proportional to the geographic distance in a city-size region where the curvature of earth is negligible. The optimization method is Adam. We initialized the learning rate as 0.001 but enabled dynamic learning rate in the training cycles based on the number of target convergence steps.

4.4.1 Implementation. We implement `CellSense` on one week of CBR data. Our model and baseline models are implemented with Keras and PyTorch libraries. We train and evaluate our design on a server with 8 Nvidia K40C GPUs. We set the learning rate as 0.001. For each GRU layer, we set the number of cells as 60 and initialize the GRU parameter with random values between -0.001 to 0.001.

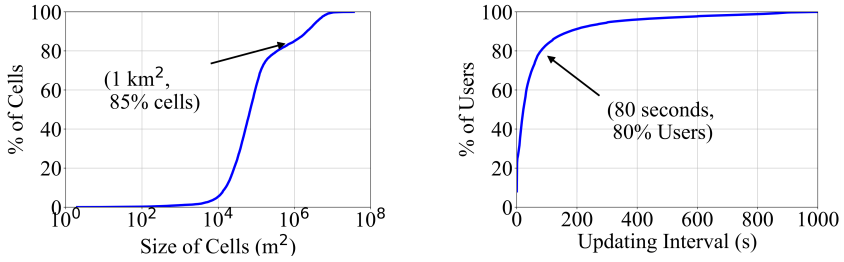
5 EVALUATION

In this section, we systematically evaluate `CellSense` with a comprehensive internal signaling data as ground truth.

5.1 Evaluation Settings

5.1.1 Signaling Data as Ground Truth. We utilize a signaling dataset as the ground truth for our model validation. Signaling data capture signal switch activities including 7 service types including Patch Switch, Circuit Switched FallBack (CSFB), Tracking Area Updating (TAU), LTE Attach, LTE Detach, LTE Paging, Service Request (e.g., 2G, 3G, 4G/LTE). When a user moves from one tower to the next tower, the device will detach from the last tower and then attach to the next cellular tower. At the same time, two records including one detach record and one attach record will be generated in the dataset. Therefore, the signaling dataset captures user locations with high

spatio-temporal coverage. If a user detached from one tower and attached to another tower, a signal switch will be captured in signaling records. When no record is found for a user in one time slot, it indicates the user stays at the same location or insides the coverage of the same towers. Therefore, we may fill users' locations at any time slot from signaling records. However, different from CBR, which are collected and are accessible for many researchers [58] [6] [40], signaling data are passively generated only for maintenance purposes during certain time periods. Fig. 12 presents spatial and temporal granularity of signaling records. Signaling data have the same spatial granularity as CBR because both data are generated on the same cellular towers. As shown in Fig. 12a, the covered cell size is less than 1 km^2 area for 85% towers. For temporal granularity, the update interval of signaling is less than 80 seconds for 80% of signaling records in Fig. 12b; whereas the update interval of CBR is less than 13 minutes for 80% of CBR users as shown in Fig. 1b. Therefore, signaling records achieve a much finer grained sensing granularity on temporal dimension compared with CBR data. Moreover, because signal level data have sector level ID that can be used to infer the more fine-grained area of users within a tower coverage, e.g., four sectors can further partition a tower into fine-grained areas, signaling record can be used as ground truth data for our mobility recovery system.



(a) Spatial Granularity (b) Temporal Granularity

Fig. 12. Granularity of Signaling Data

5.1.2 Metrics. We use the mean geographic distance, i.e., the great-circle distance, to report the estimation error between ground-truth locations and estimated locations. The great-circle distance (dist) quantifies the geographic distance between two GPS locations, e.g., 5km. We give its formula in Equation 3, where φ is latitude, λ is longitude, and R is the radius of the earth, i.e., 6,371km.

$$\begin{aligned} a &= sin^2(\Delta\varphi/2) + cos\varphi_1 \cdot cos\varphi_2 \cdot sin^2(\Delta\lambda/2) \\ C &= 2 \cdot atan^2(\sqrt{a}, \sqrt{1-a}) \\ dist &= R \cdot C \end{aligned} \tag{3}$$

5.1.3 Baselines. We implement two exiting models as our baselines, which are listed as follows.

- *HMM:* Hidden Markov Model is widely used to estimate the dependency of locations in human mobility. We adopted a widely-used HMM method to estimate the missing records. Specifically, we divided one day into 5-minute time slots and infer missing locations of users in a time slot based on the transition probability in historical records.
- *DT:* DeepTransport is a state-of-the-art method based on the LSTM method to infer detailed traces of users [49]. Different from our method, DeepTransport ignores both global, e.g., travel time, and contextual information, e.g., demographic factors. Instead, it only feeds the learning model with individual mobility records.
- *TripGen:* TripGen is one of the most recent models to infer human mobility based on CDR data. This method is built on statistical features of collective and individual mobility [2]. First, it estimates users' important

locations such as home and work locations in Voronoi partitions; second, filter valid samples and extract origin-destinations from those samples; third, integrate magnification factors in the inference.

an Ablation Study for Impact of Factors: To better understand user mobility patterns and measure the robustness the proposed model, we study the impact of three real-world factors on the performance of `CellSense`:

- The impact of the spatial dimension, e.g., performance difference in the downtown area and suburban areas;
- The impact of the temporal dimension, e.g., peak hours vs. non-peak hours;
- The impact of the population, e.g., area with high or low population density;
- The impact of the system design component, e.g., with or without collective mobility.

5.2 Evaluation Results

We report the evaluation results and the comparison with the baseline methods in terms of overall performance and the impact of listed factors.

5.2.1 Performance. As shown in Fig. 13, all three methods achieve a decent performance on the inference. When comparing `CellSense` with baseline models, we found `CellSense` outperforms both baseline models. Specifically, `CellSense` reduces the inference error from 102 meters to 39 meters on average compared with the HMM model. `CellSense` outperforms DT and reduces the inference error from 61 meters to 39 meters. Fig. 14 presents the CDF (cumulative distribution function) of inference. We found 80% of inferred locations in `CellSense` are less than 45 meters away from the ground truth, which is better than the two baseline models, i.e., 71 meters for DT, 82 meters for TripGen, and 120 meters for HMM.

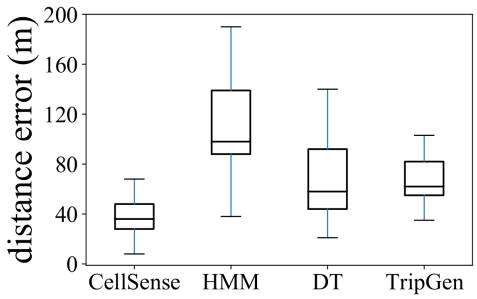


Fig. 13. Performance

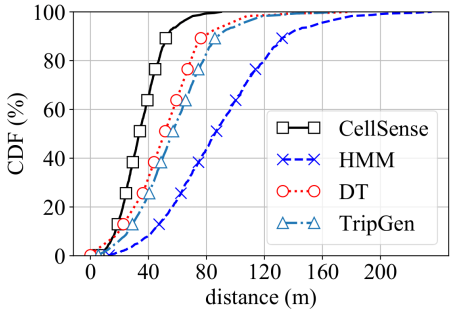


Fig. 14. Performance CDF

5.2.2 Impact of Spatial and Temporal Dimension. We further investigate the impact of the spatial dimension (i.e., different city areas) on the performance of `CellSense`. Fig. 15 illustrates the qualitative distribution of the inference performance where each point is the ground truth of users' locations, i.e., connected towers.

A red color indicates a high inference error (i.e., distance) and a yellow color indicates a low inference error. We found a lower inference error in downtown areas and major highways. Based on the analysis, we found it is caused by the tower distribution and road structures. In downtown areas, there is a denser tower distribution and users have more frequent cellular interactions. The travel distance of unobserved stages is relatively small compared with other areas. Besides, a larger amount of direct observations from users improve the accuracy of travel time estimation in downtown areas. On highways, the topological structure of the mobility graph is simple because there are always no brunches or u-turns on highways. Moreover, the variance of the travel time on highways is smaller compared with that on regular roads. As a result, we found a lower inference error on highways over other areas. On temporal dimension, we study the model performance dynamics during different hours of one day and present the result in Fig. 16. `CellSense` achieves a better performance during the day time

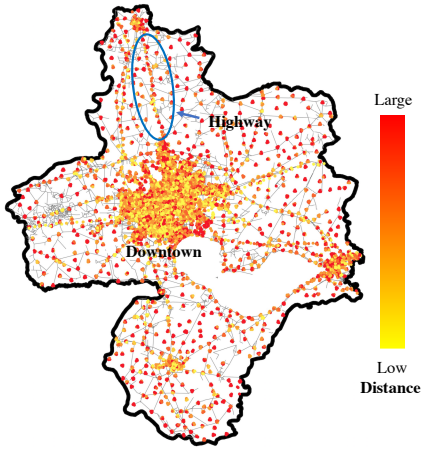


Fig. 15. Spatial Distribution

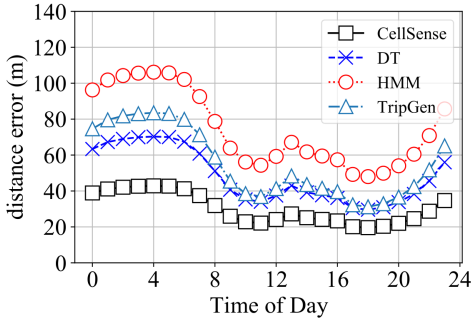


Fig. 16. Temporal Distribution

compared with the night time. We found two peaks in terms of performance around 9am and 6pm, which are two peak hours for commuters. The potential reason is two folds. First, commuting patterns are regular for most users between home and work locations. Second, users have more cellular services for internet access during home-work trips, which leads to more observations. The two factors result in a better performance of `CellSense`.

5.2.3 Impact of Population. To validate the impact of external features on model performance, we study the inference error under different population density. Fig. 17 shows the population density distribution under tower coverage. We found the population density is less than 2.4 thousand on 80% towers. In contrast, the population density is larger than 2.4 thousand per km^2 on 20% towers, which are mostly located in the downtown areas of the city. We report the detailed comparison of the performance under the impact of the population in Fig. 18, where we found the inference error increases first and then decreases with population density. When the population density is less than 3 thousand per km^2, the inference error increases with population density. However, in regions with a larger population density, the inference error decreases. Based on our analysis, we found the potential reason is that in regions with a small population density, there is a small variance of route choices and travel time, which leads to a better performance. On the other hand, in regions with a high population density, even though there is higher uncertainty of route choices and travel time for individual users, a larger amount of user records and observations are generated in those areas. As a result, it improves the model performance with more training data and direct observations.

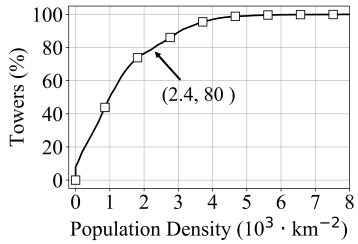


Fig. 17. Pop Density

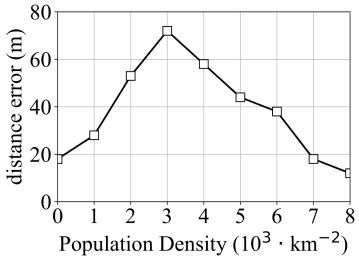


Fig. 18. Impact of Pop

5.2.4 Impact of System Design. To investigate the impact of different components in our design, we compare `CellSense` with its two variants:

- *C-P*, i.e., CellSense minus **P**ersonalized contextual information, where we drop individual contextual information from CellSense.
- *C-G*, i.e., CellSense minus **G**lobal information, where we drop collective mobility modeling from CellSense.

As shown in Fig. 19, we found both collective mobility modeling and contextual information boosts the performance of CellSense. Comparing two variant models, we found the collective mobility modeling has a higher impact on the performance of mobility recovery. In particular, the impact is larger during the commuting peak hours as shown in Fig. 20. The potential reason is that the contextual information is correlated with routine commuting patterns of users. For example, car owners prefer to drive personal vehicles for commuting purpose. Another example is the contextual factor age. A young users are more likely to have routine commuting patterns compared with elders. For collective mobility modeling, the increasing travel demand during peak hours increases the amount of observations for travel time estimation and leads to a higher estimation accuracy. Therefore, the drop of collective mobility modeling has a larger impact during peak hours compared with normal time.

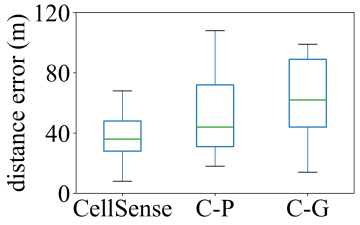


Fig. 19. Population Density

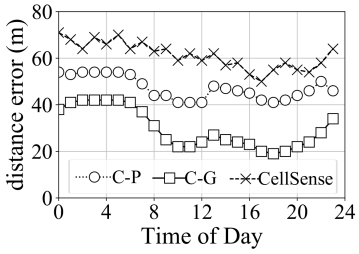


Fig. 20. Impact of Time

6 DISCUSSIONS

In this section, we share several lessons learned from our design and implementation and then discuss limitations and real-world impacts of our work.

Lessons Learned: we summarize four lessons learned as follows.

(1) **Collective Information for Data Sparsity:** Our target is to infer sensing gaps and human mobility recovery. However, CBR from an individual user are always too sparse for modeling training. Human mobility presents similar patterns under the same spatial and temporal dimension, e.g., all cars have a similar travel time and route on the same road at the same time. Therefore, to solve the data sparsity in individual users, we can infer common features without differentiating among users. The collective mobility can be used to initialize the distribution of individual mobility patterns, which can be adjusted and updated with individual mobility data.

(2) **Contextual Information for Personalization:** Apart from direct features inferred from individual traces, contextual information is an implicit feature for individual mobility. An attention mechanism or other learning strategies can be applied to capture the underlying correlations between user contextual information and mobility patterns to achieve context-aware individual mobility modeling. Such contextual information profiles users and provides insights for individual mobility modeling. Beyond mobility, we can explore many other applications with the same method such as context-aware recommendation, advertising, travel time estimation.

(3) **Sensing Stages for Irregular Sensing Granularity:** Previously, most users relied on cellular services mainly for phone calls or messages. In recent years, with the upgrades of cellular infrastructures and the high popularity of smartphones, users use cellphones more frequently for internet connections. The change of usage patterns provides new opportunities for mobility modeling from users' billing activities.

The new usage pattern enables us to divide CBR into two different stages, i.e., an observed stage with continuous CBR and an unobserved stage without observations. In the observed stage, we can infer detailed mobility status with continuous records, which can benefit the modeling on unobserved stages. The idea can be generalized into scenarios with similar irregular sensing granularity. With the same method, we can extract stages with regular sensing granularity (observed stages) and infer user features in those stages. The inferred user features can further improve the modeling of other irregular stages.

(4) **Bi-directional Mobility Inference:** Human mobility is highly correlated with both the previous and later status of the same users. Different from the prediction, in which future status remains unknown for individual users, an inference task for recovery can utilize information from both directions. The bi-directional inference used in the paper can be generalized to other sensing gap inference caused by unstable sensors or data transmissions. Moreover, a record closer to the sensing gap has a higher correlation with the missing records. Therefore, a bi-directional LSTM or GRU can be applied to automatically capture the two-dimensional correlation.

Limitations: Even a large scale of CBR and signaling data is collected and used, a major limitation of `CellSense` is we only evaluate it with the data from a single city and a single operator in China, whose users might have a bias against generic mobility patterns. We are exploring additional evaluation with cellular network data from other operators, but the internal signaling data contains detailed information of users and may rise commercial privacy concerns if we are granted access to more than one operators when those operators are always competitors to each other. Another limitation is that our work requires large-scale user records for global information inference. However, because our system is mostly beneficial to the researchers who already have CBR, this limitation may not be significant because CBR often have large-scale user records involved. Even with this limitation, we believe that the design philosophy and lessons we summarized in `CellSense`, i.e., utilizing a few real-world features to contextualize vehicular usage to better model and predict future vehicle usage, can be generalized to other scenarios.

Real World Impacts: Understanding large-scale human mobility at the individual level builds the foundation for pandemic or epidemic mitigation such as recent COVID 19. The existing study has also revealed many scenarios and applications built on human spatial-temporal information, e.g., identifying locations of crimes [12]. `CellSense` has the potential to help understanding the city resident interaction at the macro-level, which can be used to understand the coarse grained spreading path and identify potentially affected groups. Compared to the other approaches, e.g., Exposure Notification developed by Apple and Google with Bluetooth data, cellular data has two major strengths compared with other data sources for such purposes: (i) high penetration rates, i.e., most people take cellphones with them every day; (ii) low collection cost and incentive, i.e., cellular billing records are already collected for billing purpose and no extra cost or incentive is introduced for data collection. Besides, the sensing gap recovery of `CellSense` on cellular billing records provides an efficient solution to improving data quality for existing data holders such as researchers and service providers. For example, researchers with data access to either CDR (Call Detail Record) or CBR can interpolate synthetic records into the existing dataset to improve the spatiotemporal coverage of cellular data, which benefits existing models or motivates new applications. Cellular providers can provide more accurate commercial strategies with a better understanding of user distribution, e.g., building towers, discount strategies.

Ethics and Privacy: we model and infer users' sensing gap for cellular billing records with contextual demographic and behavior data, which were collected by cellular service providers under the consents of users by providing incentives such as data package rate discount. Moreover, our study is built for research purposes for social good. One important application of our model is to enable researchers to augment their sparse legacy CBR for a dense and fine-grained CBR data for their own applications [36][9][6]. Further, our inferred fine-grained

CBR data can be used (1) by cellular service providers to predict the cellular user demand [40]; (2) by public security to understand human flow prediction [31], e.g., to avoid stampedes in public events; (3) by public health to understand human mobility for epidemic spreading trace analysis and control [43]. Finally, for researchers who do not have access to real-world data, our model can be validated and improved based on artificial generation of CDRs [13]. Even with consents, we have to protect the privacy of involved users. In this project, we took the following active steps for privacy protections under IRB (Institutional Review Boards).

- Anonymization: As shown in the sample record in Tab. 2, all data analyzed is anonymized by the cellular operators, so data cannot be used to trace back to individual users;
- Secure Severs: We process all data in a secure server with access granted by cellular operators. Raw data is secured and cannot be downloaded from the secure server. All data training and evaluation are performed on the secure server, and only statistical features or evaluation results are directly reported to individual researchers.
- Minimal Exposure: We only store and process data that is useful for our sensing gap inference and human mobility modeling project, and drop other information for the minimal exposure, e.g., we drop detailed information related to users' connections. Therefore, we minimize the exposure risk of individuals in raw data. We also take active privacy protection steps to decrease the privacy leakage risks with utility preservation for our system. For instance, we decrease the frequency and normalize the weights of high-frequent locations for individual users because those locations are possible to be privacy sensitive places,e.g., home and work locations [28].
- Data Release: We only release sample data with the consent of users. For each user, only sample of data during a continuous time period is released to minimize the exposure of individual users and preserve data utility of mobility during the time period. We will normalize the location frequency from individual users, e.g., decrease data samples from high frequent locations of individual users, since high frequent locations increase user privacy concern significantly [29] [14]. We will take active protection mechanisms, e.g., differential privacy or adding synthetic records [13], to protect user privacy when we release sample user data.

Generalization: Data sparsity is a common issue in mobility modeling based on spatio-temporal datasets [6] [9] [60]. Even though `CellSense` is built on CBR data, it can be generalized to other spatio-temporal data sources describing human mobility. Specifically, it can be generalized to any other human mobility data with *location* including both exact GPS locations or relative locations, e.g., coordinates, and *time* including both exact timestamps and relative time order information. Those data sources can be either with fine-grained (e.g., continuous GPS traces, [57] [56]) or course-grained temporal gaps (discrete user-triggered check-in locations [6] [9]). In particular, to generalize `CellSense` to GPS data, we can use existing digital road maps, e.g., OpenStreetMap, as the mobility map in the implementation. For check-in data, we can construct the mobility map by combining PoI (Point-of-Interests) locations with road map topology and improve the model performance with preprocessing steps for data enhancement, e.g., route planning.

7 RELATED WORK

Due to the ubiquity of mobile devices and the unique advantage of cellular networks (i.e., high penetration rates), large-scale human location sensing based on cellular networks has been extensively investigated by peer researchers. We summarize the related work based on a two-dimension taxonomy: (i) mobility level, i.e., individual mobility or collective mobility; (ii) goal's temporal continuity, i.e., the goal is to get the temporally continuous or discontinuous mobility. The result is shown as Tabel 4.

Table 4. Mobility Sensing in Cellular Networks

Categories		Level	
		Collective	**Individual**
Goal's Temporal Continuity	**Discontinuous**	[28] [37] [30] [39] [7] [2] [32]	[20] [36] [46] [48] [47] [9] [6]
	Continuous	[64] [59] [17] [5] [45] [7]	*CellSense*

7.1 Temporally Discontinuous Mobility

Most of the existing mobility sensing work in cellular networks can be characterized in this category. Isaacman *et al.* designed *WHERE* based on Call Detail Records (CDRs) to model general human mobility based on a few important locations and evaluate it on the population movement [28]. Batran *et al.* extracted OD (origin-destination) trips from Call Detail Record (CDRs) for location attraction analyses and synethetic trip generation [2]. Jundee *et al.* inferred and visualized the users' commuting patterns from CDRs [32]. Mir *et al.* introduced *DP-WHERE* to prevent privacy leak in the *WHERE* model using differential privacy. Gonzalez *et al.* [20] modeled human mobility by studying the trajectory of 100,000 anonymized mobile phone users, and they unveiled that individuals trajectories show a high degree of temporal and spatial regularity. Pappalardo *et al.* [39] revealed two basic mobility patterns of users including exploration and preferential return individual mobility. Jiang *et al.* [30] revealed the basic patterns from CDR data to describe users' activities. Cao *et al.* [6] and Chen *et al.* [9] models and predicts user revisitation patterns on certain locations. Lin *et al.* [36] studies the correlation between mobility patterns and users' health conditions. These works are different from ours that they only focus on the mobility (i.e., individually or collectively) with discontinuous location observations, while our work focuses on inferring the missing location observations when users do not interact with cellular networks.

7.2 Temporally Continuous Mobility

Temporally continuous mobility is also studied and modeled at some works. Zhang *et al.* used multiple data sources to model human mobility at an aggregated level in real time [64]. Xu *et al.* [59] and Fang [17] *et al.* modeled the real-time collective population distribution in cities based on cellular CDR data. Cao *et al.* [7] investigated crowd mobility behaviors with cellulat tower accessing data. Calabrese *et al.* estimated the flow between origins and destinations using CDR data. These works generally focused on collective mobility such as real-time flow, while our work focused on individual-level mobility. The most similar work as ours is CTrack [50] that Thiagarajan *et al.* inferred the trajectories of individuals based on cellular tower signal fingerprinting and inertial sensors. The key difference is that it requires the participation of users (i.e., expose data from inertial sensors) that limits its scalability. In our work, all the workload is done on the cloud based on existing cellular access records and does not require extra efforts from users.

8 CONCLUSION

In this paper, we design, implement, and evaluate a human mobility recovery system named `CellSense`. `CellSense` targets to recover sensing gaps in CBR on temporal dimension. `CellSense` takes sparse CBR data as input and outputs dense continuous records to recover the sensing gap when using cellular networks as sensing systems. In `CellSense`, we design two key components: (i) an individual-independent component for collective mobility modeling; (ii) an individual-dependent component for context-aware individual mobility modeling. More importantly, we systematically evaluate `CellSense` with large-scale signaling data. The evaluation results indicate `CellSense` reduces the inference error by 35.3% compared with state-of-the-art models. We share several lessons learned from our implementation of `CellSense`. We believe the design philosophy in not restricted to human mobility recovery, and can be applied to many other real-world systems. Last but not least, Under the

consent of our collaborators, we will share one week of sample data including both cellular billing records and signaling records so peer researchers can validate and follow our work.

REFERENCES

[1] Amresh Anbalagan aand Jerry Chao and Chong Siong. 2019. China smartphone user number starts to decline - TLD by MW. https://thelowdown.momentum.asia/china-smartphone-user-number-starts-to-decline/

[2] Mohamed Batran, Mariano Gregorio Mejia, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2018. Inference of human spatiotemporal mobility in greater maputo by mobile phone big data mining. In *ATT@ IJCAI*.

[3] Filippo Maria Bianchi, Antonello Rizzi, Alireza Sadeghian, and Corrado Moiso. 2016. Identifying user habits through data mining on call data records. *Engineering Applications of Artificial Intelligence* 54 (2016), 49–61.

[4] Federica Bogo and Enoch Peserico. 2013. Optimal throughput and delay in delay-tolerant networks with ballistic mobility. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. 303–314.

[5] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing* 10, 4 (2011), 36–44.

[6] Hancheng Cao, Zhilong Chen, Fengli Xu, Yong Li, and Vassilis Kostakos. 2018. Revisitation in urban space vs. online: A comparison across pois, websites, and smartphone apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–24.

[7] Hancheng Cao, Jagan Sankaranarayanan, Jie Feng, Yong Li, and Hanan Samet. 2019. Understanding metropolitan crowd mobility via mobile cellular accessing data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 5, 2 (2019), 1–18.

[8] Guangshuo Chen. 2018. *Human Habits Investigation: from Mobility Reconstruction to Mobile Traffic Prediction.* Ph.D. Dissertation.

[9] Zhilong Chen, Hancheng Cao, Huangdong Wang, Fengli Xu, Vassilis Kostakos, and Yong Li. 2020. Will you come back/check-in again? understanding characteristics leading to urban revisitation and re-check-in. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–27.

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[11] Maria Luisa Damiani, Andrea Acquaviva, Fatima Hachem, and Matteo Rossini. 2020. Learning Behavioral Representations of Human Mobility. *arXiv preprint arXiv:2009.04719* (2020).

[12] Dorothy E Denning and Peter F MacDoran. 1996. Location-based authentication: Grounding cyberspace for better security. *Computer Fraud & Security* 1996, 2 (1996), 12–16.

[13] Viren Dias and Lasantha Fernando. [n.d.]. Generating Privacy-Preserving Artificial Call Detail Records (CDRs) From Mobile Phone Subscriber Profiles. ([n. d.]).

[14] Zhihan Fang, Boyang Fu, Zhou Qin, Fan Zhang, and Desheng Zhang. 2020. PrivateBus: Privacy Identification and Protection in Large-Scale Bus WiFi Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–23.

[15] Zhihan Fang, Guang Wang, Shuai Wang, Chaoji Zuo, Fan Zhang, and Desheng Zhang. 2020. CellRep: Usage Representativeness Modeling and Correction Based on Multiple City-Scale Cellular Networks. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 584–595. https://doi.org/10.1145/3366423.3380141

[16] Zhihan Fang, Guang Wang, and Desheng Zhang. 2020. Modeling Fine-Grained Human Mobility on Cellular Networks. In *Companion Proceedings of the Web Conference 2020*. 35–37.

[17] 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, Article 106 (Sept. 2018), 25 pages. https://doi.org/10.1145/3264916

[18] Mohammad Forghani, Farid Karimipour, and Christophe Claramunt. 2020. From cellular positioning data to trajectories: Steps towards a more accurate mobility exploration. *Transportation Research Part C: Emerging Technologies* 117 (2020), 102666.

[19] Catherine Linard Forrest R Stevens, Andrea E Gaughan and Andrew J Tatem. 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS one, 10(2)* (2015).

[20] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

[21] M. Haklay and P. Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7, 4 (2008), p.12–18.

[22] Md Mahedi Hasan and Mohammed Eunus Ali. 2017. Estimating travel time of Dhaka city from mobile phone call detail records. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. 1–11.

[23] Bilal Hussain, Qinghe Du, Sihai Zhang, Ali Imran, and Muhammad Ali Imran. 2019. Mobile edge computing-based data-driven deep learning framework for anomaly detection. *IEEE Access* 7 (2019), 137656–137667.

[24] T. Inzerilli, A. M. Vegni, A. Neri, and R. Cusani. 2008. A Location-Based Vertical Handover Algorithm for Limitation of the Ping-Pong Effect. In *Proceedings of the 2008 IEEE International Conference on Wireless Mobile Computing, Networking Communication (WIMOB '08)*. IEEE Computer Society, USA, 385–389. https://doi.org/10.1109/WiMob.2008.64

[25] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Identifying important places in people's lives from cellular network data. In *International Conference on Pervasive Computing*. Springer, 133–151.

[26] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Ranges of human mobility in Los Angeles and New York. In *2011 IEEE international conference on pervasive computing and communications workshops (PERCOM workshops)*. IEEE, 88–93.

[27] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky. [n.d.]. A Tale of Two Cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications (HotMobile '10)*.

[28] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. [n.d.]. Human Mobility Modeling at Metropolitan Scales *(MobiSys '12)*.

[29] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. 239–252.

[30] Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. 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.

[31] Wenwei Jin, Youfang Lin, Zhihao Wu, and Huaiyu Wan. 2018. Spatio-temporal recurrent convolutional networks for citywide short-term crowd flows prediction. In *Proceedings of the 2nd International Conference on Compute and Data Analysis*. 28–35.

[32] Thanisorn Jundee, Chanadda Kunyadoi, Anya Apavatjrut, Santi Phithakkitnukoon, and Zbigniew Smoreda. 2018. Inferring commuting flows using CDR data: A case study of Lisbon, Portugal. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 1041–1050.

[33] Minkyong Kim, David Kotz, and Songkuk Kim. 2006. Extracting a mobility model from real user traces. (2006).

[34] Alexey V Kurilkin, Oksana O Vyatkina, Sergey A Mityagin, and Sergey V Ivanov. 2015. Evaluation of urban mobility using surveillance cameras. *Procedia Computer Science* 66 (2015), 364–371.

[35] Yan Leng et al. 2016. *Urban computing using call detail records: mobility pattern mining, next-location prediction and location recommendation*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[36] Zongyu Lin, Shiqing Lyu, Hancheng Cao, Fengli Xu, Yuqiong Wei, Hanan Samet, and Yong Li. 2020. HealthWalks: Sensing Fine-grained Individual Health Condition via Mobility Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–26.

[37] Darakhshan J Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N Wright. 2013. Dp-where: Differentially private modeling of human mobility. In *2013 IEEE international conference on big data*. IEEE, 580–588.

[38] Published by S. O'Dea and Feb 28. 2020. Smartphone users worldwide 2020. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[39] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. 2015. Returners and explorers dichotomy in human mobility. *Nature communications* 6, 1 (2015), 1–8.

[40] Zhou Qin, Fang Cao, Yu Yang, Shuai Wang, Yunhuai Liu, Chang Tan, and Desheng Zhang. 2020. CellPred: A Behavior-Aware Scheme for Cellular Data Usage Prediction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 40 (March 2020), 24 pages. https://doi.org/10.1145/3380982

[41] 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* (Shenzhen, China) *(SenSys '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3274783.3274850

[42] Aria Rezaei, Jie Gao, Jeff M Phillips, and Csaba D Tóth. 2018. Improved bounds on information dissemination by manhattan random waypoint model. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 139–148.

[43] Stefania Rubrichi, Zbigniew Smoreda, and Mirco Musolesi. 2018. A comparison of spatial-based targeted disease mitigation strategies using mobile phone data. *EPJ Data Science* 7 (2018), 1–15.

[44] Muhammad Zubair Shafiq, Lusheng Ji, Alex X. Liu, Jeffrey Pang, Shobha Venkataraman, and Jia Wang. 2013. A First Look at Cellular Network Performance during Crowded Events. *SIGMETRICS Perform. Eval. Rev.* 41, 1 (June 2013), 17–28. https://doi.org/10.1145/2494232.2465754

[45] M Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, Shobha Venkataraman, and Jia Wang. 2016. Characterizing and optimizing cellular network performance during crowded events. *IEEE/ACM Transactions on Networking (TON)* 24, 3 (2016), 1308–1321.

[46] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (2012), 96–100.

[47] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. Modelling the scaling properties of human mobility. *Nature Physics* 6, 10 (2010), 818–823.

[48] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[49] Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. 2016. Deeptransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2618–2624.

[50] Arvind Thiagarajan, Lenin Ravindranath, Hari Balakrishnan, Samuel Madden, and Lewis Girod. 2011. Accurate, Low-energy Trajectory Mapping for Mobile Devices. In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation* (Boston, MA) *(NSDI'11)*. USENIX Association, Berkeley, CA, USA, 267–280. http://dl.acm.org/citation.cfm?id=1972457.1972485

[51] Etienne Thuillier, Laurent Moalic, Sid Lamrous, and Alexandre Caminada. 2017. Clustering weekly patterns of human mobility through mobile phone data. *IEEE Transactions on Mobile Computing* 17, 4 (2017), 817–830.

[52] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza. 2014. On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol* 10, 7 (2014), e1003716.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[54] Jinzhong Wang, Xiangjie Kong, Azizur Rahim, Feng Xia, Amr Tolba, and Zafer Al-Makhadmeh. 2017. IS2Fun: Identification of Subway Station Functions Using Massive Urban Data. *IEEE Access* 5 (2017), 27103–27113.

[55] Feng Xia, Azizur Rahim, Xiangjie Kong, Meng Wang, Yinqiong Cai, and Jinzhong Wang. 2017. Modeling and Analysis of Large-scale Urban Mobility for Green Transportation. *IEEE Transactions on Industrial Informatics* (2017), 1–1.

[56] Xiaoyang Xie, Zhihan Fang, Yang Wang, Fan Zhang, and Desheng Zhang. 2020. RISC: Resource-Constrained Urban Sensing Task Scheduling Based on Commercial Fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–20.

[57] Xiaoyang Xie, Yu Yang, Zhihan Fang, Guang Wang, Fan Zhang, Fan Zhang, Yunhuai Liu, and Desheng Zhang. 2018. CoSense: Collaborative Urban-Scale Vehicle Sensing Based on Heterogeneous Fleets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 196 (Dec. 2018), 25 pages. https://doi.org/10.1145/3287074

[58] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin. 2017. Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment. *IEEE/ACM Transactions on Networking* 25, 2 (April 2017), 1147–1161. https://doi.org/10.1109/TNET.2016.2623950

[59] Fengli Xu, Pengyu Zhang, and Yong Li. 2016. Context-aware real-time population estimation for metropolis. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1064–1075.

[60] Yu Yang, Zhihan Fang, Xiaoyang Xie, Fan Zhang, Yunhuai Liu, and Desheng Zhang. 2020. Extending Coverage of Stationary Sensing Systems with Mobile Sensing Systems for Human Mobility Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–21.

[61] Yu Yang, Xiaoyang Xie, Zhihan Fang, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. VeMo: Enabling Transparent Vehicular Mobility Modeling at Individual Levels with Full Penetration. In *The 25th Annual International Conference on Mobile Computing and Networking* (Los Cabos, Mexico) *(MobiCom '19)*. Association for Computing Machinery, New York, NY, USA, Article 11, 16 pages. https://doi.org/10.1145/3300061.3300130

[62] Yu Yang, Xiaoyang Xie, Zhihan Fang, Fan Zhang, Yang Wang, and Desheng Zhang. 2020. Vemo: Enabling transparent vehicular mobility modeling at individual levels with full penetration. *IEEE Transactions on Mobile Computing* (2020).

[63] Yu Yang, Fan Zhang, and Desheng Zhang. 2018. SharedEdge: GPS-Free Fine-Grained Travel Time Estimation in State-Level Highway Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 48 (March 2018), 26 pages. https://doi.org/10.1145/3191780

[64] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. 2014. Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking* (Maui, Hawaii, USA) *(MobiCom '14)*. Association for Computing Machinery, New York, NY, USA, 201–212. https://doi.org/10.1145/2639108.2639116

[65] Desheng Zhang, Ye Li, Fan Zhang, Mingming Lu, Yunhuai Liu, and Tian He. 2013. coRide: carpool service with a win-win fare model for large-scale taxicab networks. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. 1–14.

[66] Kai Zhao, Mohan Prasath Chinnasamy, and Sasu Tarkoma. 2015. Automatic city region analysis for urban routing. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 1136–1142.

[67] Yi Zhao, Xu Wang, Jianbo Li, Desheng Zhang, and Zheng Yang. 2019. CellTrans: Private Car or Public Transportation? Infer Users' Main Transportation Modes at Urban Scale with Cellular Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 125 (Sept. 2019), 26 pages. https://doi.org/10.1145/3351283