CellTrans: Private Car or Public Transportation? Infer Users' Main Transportation Modes at Urban Scale with Cellular Data

YI ZHAO, Tsinghua University, China
XU WANG, Tsinghua University, China
JIANBO LI, Qingdao University, China
DESHENG ZHANG, Rutgers University, USA
ZHENG YANG*, Tsinghua University, China

Understanding citizens' main transportation modes at urban scale is beneficial to a range of applications, such as urban planning, user profiling, transportation management, and precision marketing. Previous methods on mode inference are mostly focused on utilizing GPS data with high spatiotemporal granularity. However, due to high costs of GPS data collection, the previous work typically is in small scales. In contrast, the cellular data logging interactions between cellphone users and cell towers cover much higher population given the ubiquity of cellphones. Nevertheless, utilizing cellular data introduces new challenges given their low spatiotemporal granularity compared to GPS data. In this paper, we design CellTrans, a novel framework to survey users' main transportation modes (public transportation or private car) at urban scale with cellular data. CellTrans extracts various mobility features that are pertinent to users' main transportation modes and presents solutions for different application scenarios including when there are no labeled users in the studied cities. We evaluate CellTrans on two real-world large-scale cellular datasets covering 3 million users, among which 2,589 users are with labels. We assess our method not only quantitatively with labeled users, but also qualitatively with the whole population. The experiments show that CellTrans infers users' main transportation modes with accuracy over 80% (with a performance gain of 20% compared to state-of-the-art), and CellTrans remains effective when applied at urban scale to the whole population.

CCS Concepts: • **Information systems** → *Sensor networks*; **Mobile information processing systems**; • **Networks** → **Location based services**.

Additional Key Words and Phrases: cellular networks, main transportation mode, human mobility

ACM Reference Format:
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 (September 2019), 26 pages. https://doi.org/10.1145/3351283

*Zheng Yang is the corresponding author.

Authors' addresses: Yi Zhao, Tsinghua University, School of Software and BNRist, China, zhaoyi.yuan31@gmail.com; Xu Wang, Tsinghua University, School of Software and BNRist, China, darenwang11@gmail.com; Jianbo Li, Qingdao University, College of Computer Science and Technology, China, lijianboqdu@yahoo.com.cn; Desheng Zhang, Rutgers University, Department of Computer Science, USA, desheng.zhang@cs.rutgers.edu; Zheng Yang, Tsinghua University, School of Software and BNRist, China, hmilyyz@gmail.com.

2474-9567/2019/9-ART125 $15.00
https://doi.org/10.1145/3351283

1 INTRODUCTION

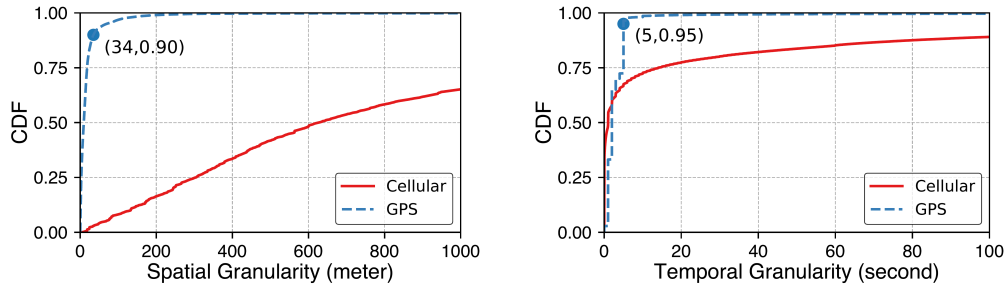
Understanding users' main transportation modes is important for various commercial and societal applications, including city planning, user profiling, transportation management, precise marketing, and location-based services. The inference of trajectory's transportation modes has been well-studied on detailed GPS and phone sensor data (*e.g.* accelerator and gyroscope) [5, 32, 49]. Nevertheless, their performance is validated by small-scale surveys, which may not be representative for urban-scale population. For example, The Geolife dataset [5, 49] contains GPS trajectories for 182 users; the mobile sensor dataset used by the Sussex-Huawei location-transportation recognition challenge [10, 37] contains only three users' data. Analyzing transportation modes at urban scale with GPS data or phone sensors is almost impossible due to the difficulty of data collection.

Luckily, the emergence of cellular networks has brought new opportunities for the urban-scale main transportation mode analysis. Cellular data can provide information about users' locations with unprecedentedly large spatial coverage in a long period (from months to years), and the application of such data takes nearly no extra cost since it is already collected for billing and monitoring purposes. These advantages have made cellular data an ideal data source for urban sensing [3, 8, 22, 42].

So, the scientific question we try to answer in this paper is *Can cellular data be used to infer users' main transportation modes?* Since previous works have designed multiple methods to infer trajectories' transportation modes utilizing various data [5, 19, 35, 49], a direct solution would be: first, identify the transportation modes of the user's all trajectory segments, and then find the most common one as the user's main transportation mode. However, this straightforward approach does not work in our scenario. Different from GPS data or cellphone sensor data, cellular data (*e.g.*, Call Detail Records CDR or Internet Access Log) cannot provide enough fine-grained information to differentiate the transportation modes of trajectories, especially for the modes that have similar speeds, *e.g.*, driving car and taking public transportation. The reasons are two fold:

- **Coarse spatial granularity:** In cellular data, the locations of connected cell towers are used to approximate the users' locations. The spatial granularity is decided by the density of cell towers, which are usually hundreds to thousands of meters away from each other. In comparison, the localization error of GPS can be smaller than 10 meters [45]. Fig. 1a compares the spatial granularity of our cellular data (detailed in Sec. 2) and Geolife GPS dataset [50, 52, 53] by the CDF of the distance between consecutive distinct location records. For Geolife GPS data, most consecutive records are less than 50 meters away; for the cellular data, the spatial granularity can be as coarse as hundreds or even thousands of meters.
- **Irregular temporal sampling:** Cellular data's generation is usually triggered by some user behaviors, such as calling or accessing Internet. So the sampling frequency is decided by how often users use their phones. In comparison, GPS is sampled regularly every few seconds. Fig. 1b compares the temporal granularity of cellular data and GPS by the CDF of inter-record intervals. Nearly all GPS intervals are shorter than ten seconds while the CDF for cellular data has a long tail towards large intervals.

To address this granularity challenge, we present a new framework in this paper to investigate users' main transportation modes at urban scale with cellular data: CellTrans. Instead of focusing on each trajectory segment, CellTrans considers a long period of users' location records, covering users' both mobile status and stationary status. The expansion of observation time can compensate for the coarse spatiotemporal granularity of cellular data and provide enough information to extract features that are pertinent to users' main transportation modes. Specifically, CellTrans extracts mobility features from three aspects: (i) movement range features that characterize users' activity space, (ii) trips' statistics that describe how often and how fast they usually travel, and (iii) user behavior features that tell us the basics of users' living patterns. Compared to velocity and acceleration features, mobility features are extracted from users' movements in a longer period and a larger area. Thus, mobility features are not sensitive to the noise and flaws of cellular data.



(a) Spatial Granularity (b) Temporal Granularity

Fig. 1. The comparison of Cellular data and GPS on (a) spatial and (b) temporal granularity. (a) is the CDF of the distance between consecutive distinct location records. (b) is the CDF of the inter-record interval. The curve of Cellular data is drawn based on our datasets and the curve of GPS is drawn based on the Geolife GPS dataset [50, 52, 53].

Based on the mobility features, we design methods to infer main transportation modes in various application scenarios: *(i) When there are labeled users* whose main transportation modes are known by app-based online survey, we show that even the number of labeled users is very limited, we can still train a supervised model that works well at urban scale on the whole population; *(ii) when there are no labeled users*, we design a clustering based method and also show the generalization capacity of the model trained on mobility features. Finally, we conduct comprehensive experiments to assess our methods' performance at urban scale with two cellular datasets that cover 3 million users within 48 days.

The main contributions of the work are summarized below:

- We design a new framework to survey users' main transportation modes (public transportation or private car) at urban scale with coarse-grained cellular data. Instead of focusing on each trajectory segment, we reveal the relationship between users' general mobility features and their main transportation modes. CellTrans extracts robust features that are insensitive to cellular data's low spatiotemporal granularity and takes advantage of cellular data's event-driven property to characterize users' behavior patterns.
- CellTrans can be applied in various application scenarios with or without labeled data. For scenarios where labeled data exist and can be used to train a model, we design a set of supervised methods given features we extracted. More importantly, considering the difficulty to collect labeled users' data for model training, we design unsupervised methods that work well without any labeled users from the studied city, including a clustering-based method and transferred models from other cities. This makes our solutions more practical and easier to apply to other cities.
- We evaluate the accuracy of CellTrans on two large-scale datasets from two cities. One city's dataset records 1,835,509 users' Internet access behaviors in 48 days, and the other city's dataset covers 1,077,106 users within 48 days. Among all these users, 2,589 users are with labels to indicate their main transportation modes, which are used for training and evaluation. By the quantitative evaluation, we show that our method's accuracy is higher than the state-of-the-art by 20%, and even when there is no training data at all, we can still achieve an accuracy about 80%. By the qualitative evaluation, we show that our methods remain effective when applied to the whole population at urban scale.

In the following sections, we introduce some terms and our datasets in Sec. 2 and present the overall framework of CellTrans in Sec. 3.1. Then we introduce the three modules of CellTrans in Sec. 3.2, Sec. 4, and Sec. 5. The

Table 1. Example of mobile flow records.

User	Time	Tower	Tower location	HTTP host	HTTP URI
xxxx	2016-12-20 16:40:01	572636	(xxx, xxx)	m5.amap.com	ws/mapapi/navigation/auto
...	...	...	...	...	...

Table 2. Shenyang Dataset

Statistics	Value
Records	8×10^9
Cell towers	1.2×10^4
Covered users	1.8×10^6
Covered area	1.3×10^4 km^2
Covered period	Dec. 19, 2016 - Feb. 4, 2017

Table 3. Dalian Dataset

Statistics	Value
Records	12×10^9
Cell towers	1.2×10^4
Covered users	1.1×10^6
Covered area	1.3×10^4 km^2
Covered period	Dec. 19, 2016 - Feb. 4, 2017

performance is evaluated in Sec. 6. The discussion and related works are in Sec. 7 and Sec. 8. Finally, we conclude the paper in Sec. 9.

2 PRELIMINARY

2.1 Terminology

We explain some terms in this section, including *mobile flow record*, *trajectory*, *stay*, and *trip*.

- **Mobile flow record:** Mobile flow records (MFR) are system logs of users' Internet access behavior in cellular networks. When there is a packet transmitted between a mobile phone and its connected cell tower, the monitoring system of the cellular network generates a mobile flow record. The record notes down the user's ID, time, tower's ID of this connection, as well as the HTTP host and URI of the request (see Table 1).
- **Trajectory:** Trajectory is a user's chronologically sorted location records: $< t_1, p_1 >, ..., < t_n, p_n >$, where t_i and p_i are the time and location of this user's ith mobile flow record.
- **Stay:** A stay is that a user remains in a geographic neighborhood longer than a time threshold. Stays are usually related to users' activities like resting at home, working at office, and hanging around in a park.
- **Trip:** A trip is a continuous subsequence of a user's trajectory between two neighboring stays. It is after a user leaves his/her last stay point and before he/she arrives at the next stay point.

2.2 Dataset

We base our design on two large-scale MFR datasets from different cities: Shenyang and Dalian. The two cities are both located in the northeast of China. The datasets cover 3 million anonymized mobile phone users in 48 days. The basic statistics of the datasets are shown in Table 2 and Table 3. Also, the individual, spatial, and temporal distribution of MFR is shown in Fig. 2.

The average number of records per person per day is 91 in Shenyang and 226 in Dalian, but the distribution among users are uneven as shown in Fig. 2a, so is the distribution among cellular towers (Fig. 2b). Fig. 2c shows the variation of the number of records with time. Before dawn, MFRs are generated slowly because most users are in sleep. At night, the generation of MFRs reaches its peak.

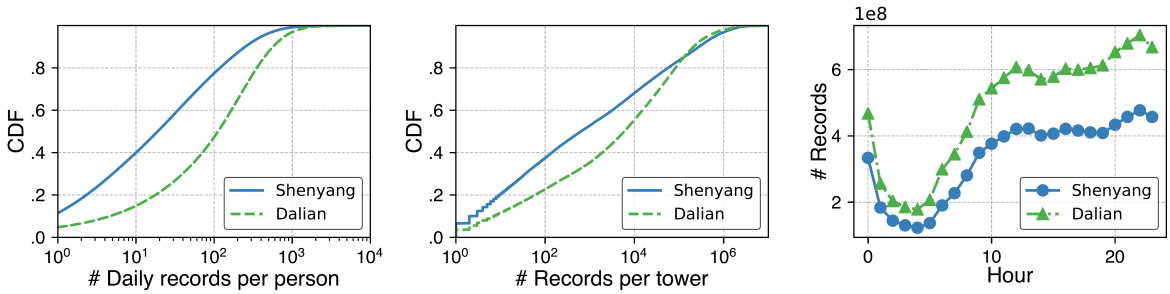


Fig. 2. (a) Individual, (b) spatial, and (c) temporal distribution of MFR. (a) CDF of the number of daily records per person. (b) CDF of the number of records per cellular tower. (c) The number of records generated in each hour in a day.

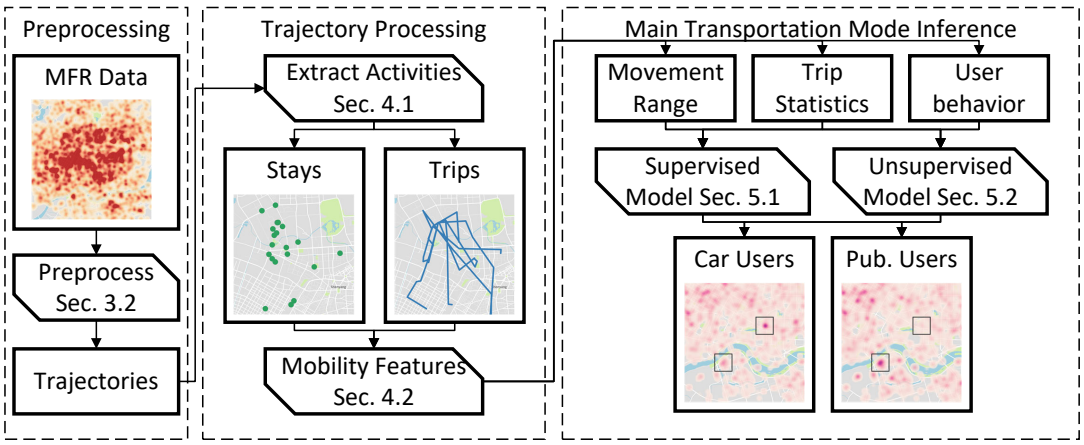


Fig. 3. Framework of CellTrans. Pub. stands for public transportation.

3 OVERVIEW AND PREPROCESSING

This section presents the overall framework of CellTrans (Sec. 3.1) and the preprocessing of MFR (Sec. 3.2).

3.1 Overview

CellTrans consists of three modules (see Fig. 3):

- **Preprocessing:** This module preprocesses the noisy MFR data (Sec. 3.2). It deals with the oscillation [28] (ping-pong effect) and bursty sampling problems.
- **Trajectory Processing:** This module first extracts users' stays and trips from the rectified MFR trajectories, and then extracts mobility features based on users' stays and trips (Sec. 4). The mobility features are from three categories, *i.e.*, movement range, trip statistics, and user behavior features.
- **Main Transportation Mode Inference:** This module infers users' main transportation modes (Sec. 5) based on the mobility features. We consider two application scenarios: when we have some labeled users whose main transportation modes are known, we can train a supervised model on these users; when we do

not have any labeled users, we design a clustering-based method and investigate the transferability of the models trained on other cities' labeled users.

3.2 Preprocessing

The preprocessing of MFR needs to deal with two problems:

Oscillation: Oscillation problem (ping-pong effect) occurs when mobile phones switch between towers rapidly, even though users do not move at all [28]. In this paper, we utilize the method designed by Yang et al. to deal with the oscillation problem. First, a user's mobile flow records are sorted chronologically as $< t_1, p_1 >, < t_2, p_2 >, ..., < t_i, p_i >, ...$ where t_i is the timestamp and p_i is the location of the i^{th} record, then a record $< t_i, p_i >$ is filtered out as oscillation noise if:

$$\begin{aligned} &Dist(p_{i-1}, p_i) > d_{thresh} \quad \& \quad Dist(p_i, p_{i+1}) > d_{thresh} \quad \& \\ &t_i - t_{i-1} < t_{thresh} \quad \& \quad t_{i+1} - t_i < t_{thresh} \end{aligned} \tag{1}$$

where $Dist$ is the Euclidean distance function, d_{thresh} is a threshold for long distance, and t_{thresh} is a threshold for short time. d_{thresh} is set to 3.6 km and t_{thresh} is set to 108 s following the method in [48].

Bursty Sampling: The generation of MFRs used in this paper is triggered by users' Internet access behaviors. Based on some previous studies, people usually access Internet resources in a bursty manner [16, 38]. When people use their phones intensively, many records are generated in a short time, which causes redundancy in the location data. We take a record $< t_i, p_i >$ as redundancy and filter it out if $t_i - t_{i-1} < 10$ s and $p_i = p_{i-1}$.

4 TRAJECTORY PROCESSING

In this section, we introduce the trajectory processing module. This module includes two parts: first, we identify users' stays and trips from the preprocessed trajectories (Sec. 4.1); then we extract the mobility features based on users' stays and trips (Sec. 4.2).

4.1 Stays and Trips

Parsing users' raw cellular data into stays and trips is the basis for further analysis [1, 6]. Stays usually correspond to users' activities like resting at home or working at office. Trips are trajectory segments when users travel from one stay region to another by some transportation means. In this paper, we use the method presented by Jiang et al. to extract stays from users' cellular data and then, we identify the trajectory segments between adjacent stays as trips.

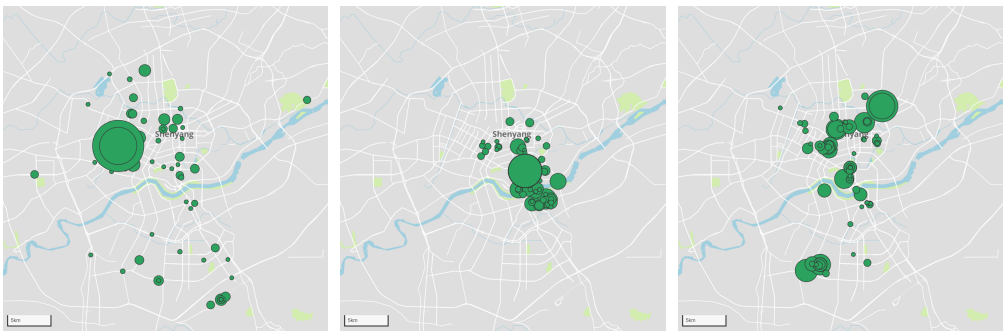
4.2 Mobility Features

CellTrans extracts mobility features from three categories:

4.2.1 Movement Range. Features in this category reveal the range of a user's movement. Compared to public transportation, driving car is a more convenient and flexible transportation means. For users whose main transportation modes are driving car, it is easier to travel to faraway places. So intuitively, the movement range of car users tend to be larger than that of public transportation users. For instance, Fig. 4a and Fig. 4b are a car user and a public transportation user, and the size of their movement range is quite different. The first feature to describe a user's movement range is:

- **Radius of gyration** (r_g): The radius of gyration has been widely adopted to characterize the mobility level of a user [11, 29]. It is calculated based on a user's stays:

$$r_g = \frac{\sum_{i=1}^{n} Dist(s_i.p, cm) * s_i.t}{\sum_{i=1}^{n} s_i.t} \quad cm = \frac{\sum_{i=1}^{n} s_i.p * s_i.t}{\sum_{i=1}^{n} s_i.t} \tag{2}$$



(a) A car user. (b) A public transportation user. (c) A public transportation user.

Fig. 4. Visualization of 3 users' stays: one car user (a) and two public transportation user (b)(c). The green circles represent users' stays. The radius of the circle corresponds to the stay time. Larger circle means that users spend more time there. (a) $r_g = 5591\,\text{m}, n_{cluster} = 39, a_{convex} = 523\,\text{km}^2$, (b) $r_g = 2058\,\text{m}, n_{cluster} = 12, a_{convex} = 52\,\text{km}^2$, (c) $r_g = 7204\,\text{m}, n_{cluster} = 22, a_{convex} = 179\,\text{km}^2$

where r_g is the user's radius of gyration, n is the number of the user's stays, $s_i.p$ and $s_i.t$ are the position and interval of the i^{th} stay, $Dist$ is the Euclidean distance function, and cm is the center of mass of the user's stays.

There is a drawback with r_g. It can only reflect one-dimension information of the distribution of users' stays. For example, if there are n points distributed evenly on the edge of a circle. No matter how n varies, the r_g of them will always equal to the circle's radius. In order to measure users' movement range comprehensively, we consider two other features:

- **Number of distinct stay clusters ($n_{cluster}$)**: We cluster the stays of a user with *DBSCAN* [7], a widely used spatial clustering method [43, 44]. The number of clusters can disclose how many distinct places a user visits. We do not use the number of stays directly because some stays may be very close to each other, and they are visits to the same place at different times. In our case, the places which users only visit one or two times are also important. In order to capture these outlier places, we set the threshold *min_samples* of DBSCAN to 1. This threshold controls how many stays at least can form a cluster. With this setting, even a stand-alone stay is far from all the other stays, it will form a cluster, too.
- **Convex hull area of stays (a_{area})**: The area of the stays' convex hull can be an indicator of how much area a user's activities cover.

A real-life example can be seen in Fig. 4, the public transportation user in (c) has a very large r_g, even larger than the car user in (a). But if we take $n_{cluster}$ and a_{area} into consideration, (c) actually has a smaller movement range than (a). Therefore, together with the radius of gyration, the number of stay clusters and area of convex hull provide comprehensive characterization of a user's movement range.

4.2.2 Trip Statistics. The features in this category characterize users' trips. Due to the low spatiotemporal granularity, it is nearly impossible to calculate the real-time speed of trips. However, for the inference of users' main transportation modes, the high-level statistics of trips can provide useful information:

- **Number of trips (n_{trip})**: As discussed in Sec. 4.2.1, having car reduces the time cost for users to travel to other places. Thus if two users have the same level of desire to travel, the user that have car tends to go out more often, thus have more trips than the user whose main transportation mode is public transportation.

Table 4. Summary of Mobility Features

	Feature	Description
Movement Range		
	r_g	Radius of gyration
	$n_{clusters}$	Number of distinct stay clusters
	a_{area}	Convex hull area of stays
Trip Statistics		
	n_{trip}	Number of trips
	n_{nt}	Number of night trips
	av_n_{trip}	Average number of trips per day
	av_n_{nt}	Average number of night trips per day
	av_v	Average speed of trips
	av_t	Average time of trips
User Behavior		
	i_{net}	Network access intensity
	t_{leave}	Time to leave home
	t_{back}	Time to come back home
	p_{house}	House price (income level)

- **Number of night trips (n_{nt})**: This feature is based on the common sense that the public transportation system is mostly not in operation at night (usually after 11:00 pm). So the high-speed trips at night are presumably car trips since no buses are running. This feature is helpful to find car users who have rich night activities.
- **Average number of (night) trips per day (av_n_{trip} and av_n_{nt})**: We notice that the days that users have records are different for different users. Some users appear in more days than others. To compensate for the sampling bias, we divide the above two features by the number of days when users have records.
- **Average speed and average time of trips (av_v and av_t)**: For a public transportation trip, users usually need to walk to the bus station first, and after getting off the bus, they often have to walk again to their final destination. Besides, the bus stops at intermediate stations instead of heading directly to the destination. As a consequence, same-length bus trips are usually slower and more time-consuming than car trips.

4.2.3 User Behavior. We consider the features related to users' daily behaviors and economical status. Cellular data's event-driven property causes the uneven sampling and other problems. Being event-driven can also be an advantage of cellular data: it enables the cellular data to contain user behavior information, *e.g.* how heavily the user uses his/her phone and when the user begins to use the phone actively in the morning. In the third category, we take this advantage and consider the features that can describe users' daily behaviors, including their network access behaviors and daily schedules.

Mode preference is also influenced by users' economical status. Wang et al. take users' socioeconomic features into the inference of transportation modes, including whether having a bus card, the number of household bicycles, and the number of household cars. However, these information is difficult to collect at urban scale. In this paper, we use the house price of the area where people live as a feature to reflect users' income level, which is easier to acquire at large scale with low cost.

- **Network access intensity (i_{net})**: MFR records users' locations in case they retrieve data from cell towers. This event-driven mechanism makes MFR capable of approximating users' phone usage behaviors. Users

driving cars use their phones less frequently than the users on buses because they need to focus on driving. We calculate the average frequency at which users generate records during trips as an indicator of the network access intensity (i_{net}).

- **Time to leave home (t_{leave}) and come back home (t_{back})**: We utilize the average time when users leave and come back home to describe their basic life patterns. We identify users' homes as the places they spend most time at night (from 20:00 to 08:00 the next day). This simple heuristic rule has been applied in a few previous works to identify places of residence [17, 26, 43].
- **House price (p_{house})**: The house price can reflect users' income level. People with low income usually prefer public transportation. The house price is acquired from [40].

We summarize the mobility features introduced in this section in Table 4. The importance and distribution of the features will be discussed in Sec. 6.4.

5 MAIN TRANSPORTATION MODE INFERENCE

Based on the mobility features extracted in the last section (Table 4), we can infer a user's main transportation mode. In some scenarios, the operators or the governments have cellular data and some labeled users whose main transportation modes are already known. In this case, a supervised model can be trained (Sec. 5.1). In other scenarios, there are no labeled users whose transportation modes are known. For such scenarios, we design unsupervised methods that does not have a training process (Sec. 5.2).

5.1 Scenario 1: With Label Data

In the first application scenario, we assume that partial users' actual main transportation modes are known. So a model can be trained to relate citizens' mobility features to their main transportation modes. Multiple supervised learning models can be applied in this scenario, *e.g.*, Decision Tree, Support Vector Machine and Multilayer Perceptron, and their performance will be evaluated in Sec. 6.2.1.

Fig. 5 shows how CellTrans infers users' main transportation modes in this scenario: first, we find out the users whose main transportation modes are very likely on foot by rules. These users are supposed to have a relatively smaller movement range compared to car and public transportation users. So a user is identified as travelling mainly on foot if his/her r_g is smaller than a threshold r_{thrd}.

We acknowledge that such a method may not be robust. There can be users who prefer driving or taking taxis for very small distances. These users may be wrongly taken as traveling-mainly-on-foot users. Actually, MFRs record locations at the granularity of cell towers, so they can not characterize users' movement in short ranges. Also, we donot have groundtruth for traveling-mainly-on-foot users to evaluate the accuracy. So here we choose a relatively simple method to identify major on-foot users and focus on the differentiation between private car and public transportation.

Among the remaining users, we train a supervised model with the labeled users to infer users' main transportation modes between private car and public transportation based on users' mobility features. Finally, we apply the model to all the users whose r_g is larger than r_{thrd} to classify them as car users or public transportation users. The determination of the threshold will be discussed in Sec. 6.2.2.

5.2 Scenario 2: Without Label Data

The training of the supervised model in Sec. 5.1 needs labeled data. However, it is not trivial to label users' transportation modes. So in this section, we introduce the methods to survey users' modes when there are no label data in the studied cities. First, we introduce a clustering-based method in Sec. 5.2.1. Then we discuss the transferability of the supervised models between cities in Sec. 5.2.2.

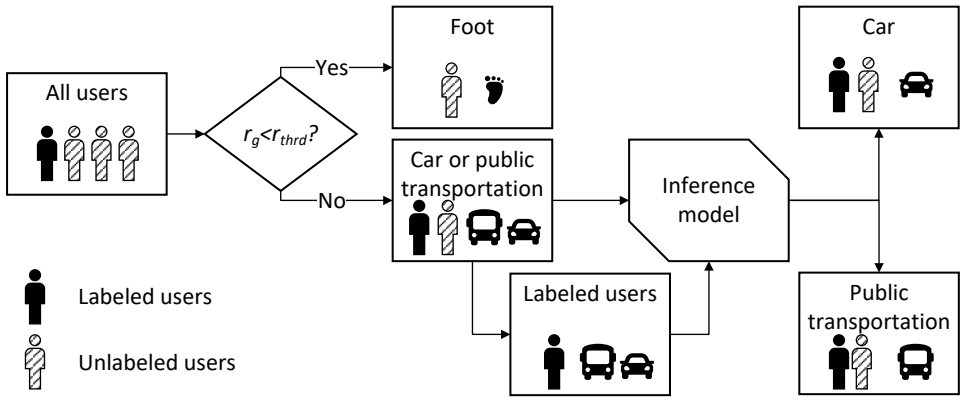


Fig. 5. Framework of the method for scenario 1. First, on-foot users are identified based on r_g. Among the remaining users, we train a supervised model to differentiate car users and public transportation users.

5.2.1 Clustering-based Method. This method first clusters the users into groups using the mobility features. We assume that the users in the same group have the same mode. To find out the modes of the groups, we show domain experts the group centers' mobility features and ask them to infer each group as "private car" or "public transportation". For example, if a group center's $n_{cluster}$ is larger than others, this group's users are probably car users. After the group centers' modes are inferred, all the users are inferred to have the mode of their group centers.

To evaluate the performance of this method independently of experts' inference, we study the upper bound of this method's accuracy as following:

Assume U to be the set of total users, $U_l \subseteq U$ to be the set of labeled users. We do not use these labeled users to train the model, instead, we use them to assess the method's performance. After clustering, we get n_c clusters $C_0, C_1, ..., C_{n_c-1}$, where C_i is the set of users in the i^{th} cluster. For each cluster C_i, $U_l \cap C_i$ is the labeled users in this cluster. We take the most common mode in $U_l \cap C_i$ as the inferred mode for this cluster. Then all users in C_i are inferred to have this mode. It is evident that this is the best inference an expert can make and the accuracy of this inference will be the upper bound of the clustering-based method's accuracy.

The clustering method we adopted is *k-means*. The selection of the parameter k of *k-means* is a tradeoff: if k is too large, the upper bound of accuracy will be high. For example, if k equals the number of users ($k = |U|$), although the upper bound will be 100%, the expert's inference of clusters' modes would be very difficult, and the clustering would be meaningless. On the other hand, if k is too small, the needed expert efforts will be minor, but the accuracy is not likely to be high. The proper value of k will be studied in Sec. 6.3.

5.2.2 Transfer Models between Different Cities. In this section, we discuss the transferability of the supervised models in Sec. 5.1. Assume that there are two cities: city A and city B. In city A, we have labeled users whose transportation modes are known. In city B, there are no labeled users. The question is: can we use the model trained on city A's labeled users to infer the transportation modes of city B's users? There are two reasons why the supervised model of CellTrans is transferable between cities:

- We extract general features about users' mobility patterns without directly utilizing the low-level city-dependent features like the longitude and latitude of users. High-level features are usually more generalizable than low-level features.

Table 5. Shenyang Groundtruth

Transportation mode	# Groundtruth users
Car	679
Public transportation	633

Table 6. Dalian Groundtruth

Transportation mode	# Groundtruth users
Car	813
Public transportation	464

- All the features are normalized. For each feature, its value is normalized into $[0, 1]$ as in Eq.(3), where $f_{ij}(f_{ij}^{norm})$ is the raw(normalized) value of the i^{th} user's j^{th} feature and U is the set of all users. Different cities may have different magnitudes, and this may influence the absolute values of mobility features. For example, in a larger city, the movement range of users may all tends to be larger. After normalization, features are transformed into relative value compared to other citizens. The normalization can decouple the city-dependency from the mobility features, thus improving the model's transferability.

$$f_{ij}^{norm} = \frac{f_{ij} - min(\{f_{ij}|i \in U\})}{max(\{f_{ij}|i \in U\}) - min(\{f_{ij}|i \in U\})} \tag{3}$$

To summarize, if we want to infer users' modes in a city where there are no labeled users, we have two options: (i) the clustering based method in Sec. 5.2.1. (ii) the transferred supervised model from other cities (Sec. 5.2.2).

6 EVALUATION

First, we introduce the evaluation methodology in Sec. 6.1. Then we evaluate the performance of CellTrans under different application scenarios in Sec. 6.2 and Sec. 6.3. We analyze the importance and distribution of mobility features in Sec. 6.4. Finally we explore the impacts of spatiotemporal granularity in Sec. 6.5

6.1 Evaluation Methodology

Datasets: We evaluate the performance with two large scale MFR datasets from two cities in the northeast of China: Shenyang and Dalian. The statistics about the two datasets are introduced in Sec. 2.2.
Inference task: The task is to infer users' main transportation modes based on their MFRs. We first extract users' mobility features as in Sec. 4.2 from MFR, and then apply supervised or unsupervised models to infer the main transportation modes.
Ground truth: The groundtruth of users' main transportation modes is from the usage analysis of Amap, one of the most popular navigation applications in China. We find that when users use Amap on their phones, different transportation modes' navigation will generate different HTTP URIs.

We count the number of days that users use either mode's navigation service. Users with more than 17 days' usage of car navigation and no public transportation navigation are identified as groundtruth car users; users with more than three days' usage of public transportation navigation and no car navigation are identified as groundtruth public transportation users. The reason that the two thresholds are different is: people are more likely to use navigation applications when they drive cars than take public transportation. The numbers of groundtruth users in two cities are listed in Table 5 and Table 6.
Performance metrics: We use three metrics to assess the performance of different methods: accuracy, precision, and recall. Precision and recall are averaged over the two classes.
Baseline: Previous works have designed multiple methodologies to infer the transportation mode of a trip. As baselines for our inference task, we first infer trips' transportation modes with these methods and then the most common transportation mode of a user's trips is considered to be the main mode. The baselines are:

Table 7. Methods to infer main transportation mode.

Method	Input
CellTrans with SVM	User mobility features Sec. 4.2
CellTrans with Decision Tree (Tree)	User mobility features Sec. 4.2
CellTrans with Random Forest (RF)	User mobility features Sec. 4.2
CellTrans with MultiLayer Perceptron (MLP)	User mobility features Sec. 4.2
Baseline 1: Aggregation with Decision Tree (Agg. Tree)	Trip velocity/acceleration features [49]
Baseline 2: Aggregation with CNN (Agg. CNN)	Trip velocity/acceleration sequences [5]

- **Agg. Tree:** Zheng et al. extract basic and advanced features from trajectories including the velocity, acceleration, heading direction and so on. A decision tree is applied to infer trips' transportation modes based on these features.
- **Agg. CNN:** Dabiri and Heaslip devise a *CNN* based model to infer GPS trips' transportation modes. It uses a one-dimension convolution network on the sequences of speed, acceleration, jerk, and bearing rate.
- **Agg. Station:** For cellular data, Li et al. suggest identifying bus trips by measuring the closeness between trips' starting and end points to the bus stations.
- **Agg. Cluster:** Wang et al. present another method focusing on cellular data. They first aggregate the trips that have close starting points and end points. Then each group of trips are clustered into two clusters by their traveling time. The group with shorter traveling time is identified as car trips, and the other group is public transportation trips.
- **Route matching:** Phithakkitnukoon et al. design a method to determine users' transportation modes based on cellular data. They query the driving and public-transport routes using map service API. Then they measure the distance between the routes and users' records. The mode of the closest route is taken as the user's mode.

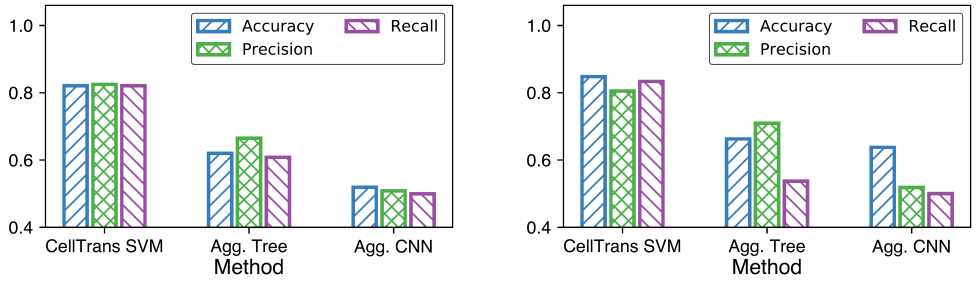
Due to the lack of ground truth, the last three baselines focusing on cellular data have not been evaluated at individual level in their original papers. In the following experiments, we take the first two methods as the baselines for our first application scenario (Sec. 6.2) because they need labeled data for their training process; the last three methods, which are rule-based and do not have a training process, are the baselines for our second application scenario (Sec. 6.3).

6.2 Scenario 1: With Label Data

In the first application scenario, we assume that the operators or city governments have cellular data along with some labeled users whose main transportation modes are already known. We do not have the ground truth transportation modes for all individuals in the city, so the quantitative analysis of our method's performance is only on the labeled users (Sec. 6.2.1). The performance on the whole population is evaluated qualitatively (Sec. 6.2.2).

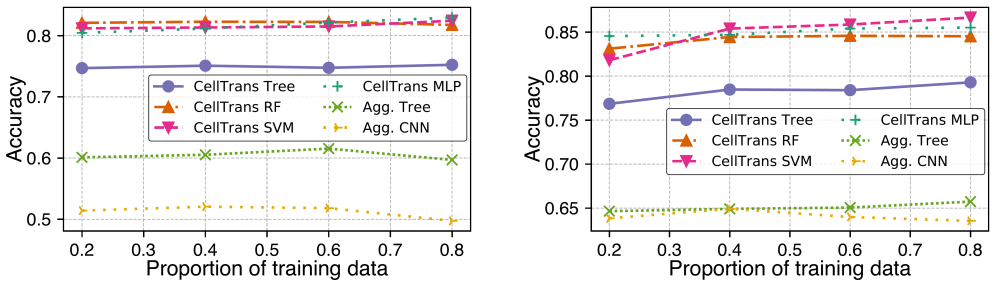
6.2.1 Quantitative Evaluation. In this section, we compare the performance of our method introduced in Sec. 5.1 with previous methods. The comparing methods are summarized in Table 7. The SVM model uses radial basis function kernel. We evaluate different methods with 5-fold cross-validation.

As shown in Fig. 6, our method that applies SVM to mobility features performs consistently better than the baselines through different metrics and different cities. In Shenyang, our method improves the accuracy by 20%, precision by 16% and recall by 21%; in Dalian, our method improves the accuracy by 19%, precision by 10% and recall by 30%. The low spatiotemporal granularity makes it nearly impossible to extract detail velocity information



(a) Shenyang (b) Dalian

Fig. 6. Comparison of CellTrans with previous methods. (a) shows the results of Shenyang and (b) shows the results of Dalian.



(a) Shenyang (b) Dalian

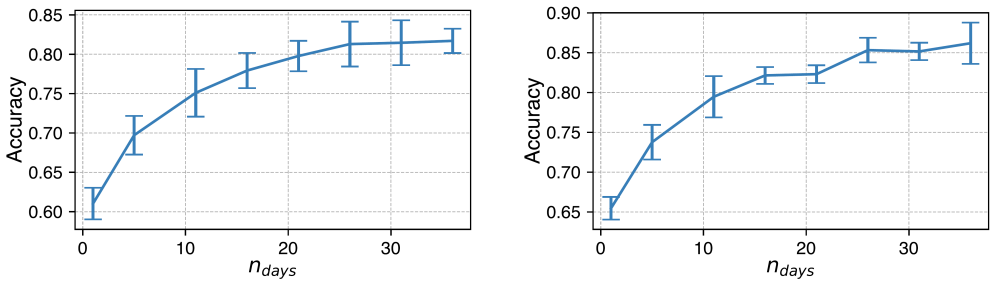
Fig. 7. Impact of the proportion of training data on different methods' accuracy. (a) shows the results of Shenyang and (b) shows the results of Dalian.

from cellular data. Therefore, previous methods depending on velocity and acceleration to infer transportation modes become ineffective.

Specifically for the CNN model, the reasons why it does not perform well are: (i) The CNN model designed by Dabiri and Heaslip takes the sequences of speed, acceleration and other features as input. However, such fine-grained features can not be extracted from cellular data accurately. (ii) Cellular data are much noisier and more irregular than GPS. It needs heavy processing to extract informative features from cellular data. CNNs can learn effective features automatically from raw data, but this is based on huge amount of training data.

Further, we try to answer two questions about the inference of users' main transportation modes:

How many labeled users do we need? The labeled users only count a small part of the city's whole population. So can we train a model with a small number of labeled users that can work well on a large number of users? To answer this question, we vary the number of labeled users that are used to train the model and observe how the accuracy evolves on the remaining labeled users. Each experiment is repeated ten times and the average accuracy is shown in Fig. 7. First, our methods outperform the baselines consistently in different cities and with different training set sizes. Second, among the methods based on users' mobility features, SVM,



(a) Shenyang (b) Dalian

Fig. 8. Impact of the number of observation days (n_{days}) on the methods' accuracy. The y-axis error bar is the standard deviation of 5-fold cross-validation. (a) shows the results of Shenyang and (b) shows the results of Dalian.

Table 8. The Proportion of Car Users Based on Census data and CellTrans.

City	Population	Cars	$\frac{\text{Cars}}{\text{Population}}$	Total Users	Stable Users	Car Users	$\frac{\text{Car Users}}{\text{Stable Users}}$	Error
Shenyang	8,292,000	1,567,000	0.19	1,835,509	321,699	31,818	0.10	0.09
Dalian	5,956,300	1,520,000	0.26	1,077,106	562,277	178,434	0.32	0.06

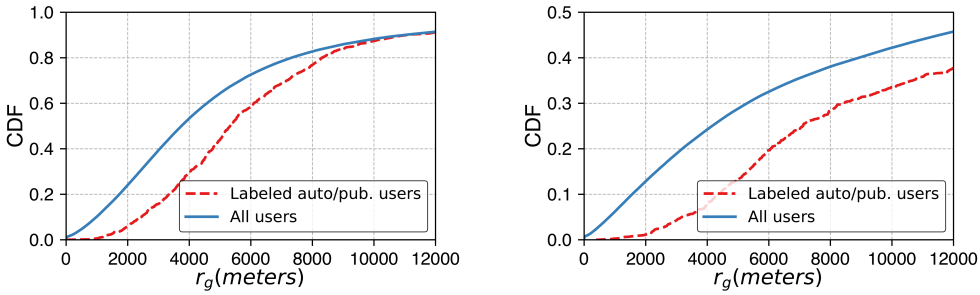
MLP and Random Forest have close performance and they all consistently perform better than Decision Tree. Third, with the increment of training data, the performance of our methods improves slightly. Even with only one-fifth data as the training set, our method still achieves an accuracy above 80%. This implies that the accuracy is not sensitive to the size of the training set. Although we only have a small number of labeled users, it should be enough to train a model that can achieve high accuracy on the whole population.

How many days of records do we need? Main transportation mode is an aspect of users' living habits. Inferring it may need a long time observation of users' activities. In this experiment, we scissor the first n_{days} of users' records to infer their main modes with SVM. Fig. 8 demonstrates how the accuracy varies with n_{days} in two cities. The y-axis error bar is the standard deviation of 5-fold cross-validation. We can see that the accuracy grows higher with a longer observation time in both cities, but the growing speed slows down gradually. To achieve an accuracy above 80%, we need more than 21 days of users' records.

6.2.2 Qualitative Evaluation. In the second part, we analyze the performance of CellTrans qualitatively at urban scale. The evaluation is from two aspects: first, we compare the statistics extracted from our results with that published by governments; second, we visualize the spatial distribution of car/public transportation users.

Before we apply the trained model to the whole population, we need to filter out the users who do not have enough data and the users whose main transportation modes are on foot. We only keep the users whose location records span more than 21 days (as we discussed previously), and we refer to these users as *stable users*.

As we mentioned in Sec. 5.1, a user is filtered out as traveling mainly on foot if his/her r_g is smaller than a threshold (r_{thrd}). Because we do not have groundtruth for on-foot users, we can not set the threshold r_{thrd} directly based on the precision and recall of identified on-foot users. We set the threshold instead by analyzing how many known private car or public transport users are mistaken as on-foot users. Fig. 9 shows the cumulative distribution function of r_g of labeled car/public transportation users and all users. Based on Fig. 9, we set $r_{thrd} = 2000$ m. We



(a) Shenyang (b) Dalian

Fig. 9. CDF of r_g of labeled auto/public transportation users and all users. (a) shows the results of Shenyang and (b) shows the results of Dalian.

can see that with such r_{thrd}, few users with known labels (private car or public transport) are taken as on-foot users, compared to the relatively large proportion in all users. We acknowledge that this can only indicate the accuracy of identified on-foot users in a very limited and indirect way, but it is the best way we can think of in this situation.

Comparison with Census Data: After the user filtering, we feed the rest users into the trained model to decide their main transportation modes. To validate the method's performance at urban scale, we estimate the proportion of car users in the whole population. The results are compared with the cities' census data in Table 8. The second to fourth columns are from census data published by the cities' governments [23–25]. The fifth to eighth columns are from our datasets and the inferred results. The last column is the error of the estimated proportion of car users, which is 0.09 in Shenyang and 0.06 in Dalian. To some extent, the results indicate the effectiveness of our method on the whole population.

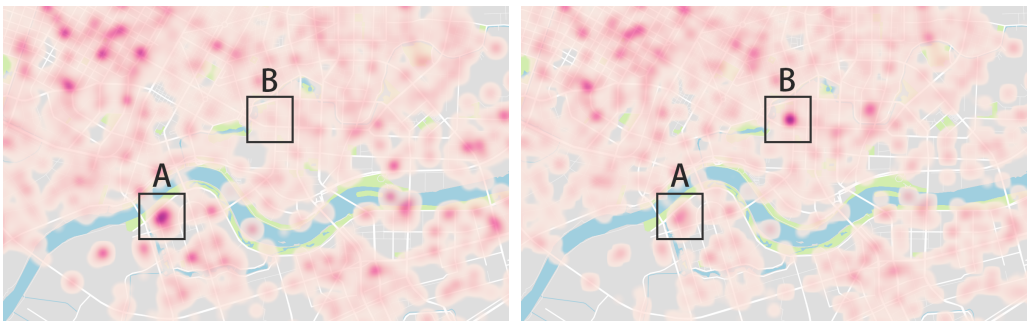
Spatial Distribution: In Sec. 4.2, we identify users' homes. In this experiment, we visualize the distribution of car/public transportation users' homes in the form of heatmap (Fig. 10 and Fig. 11), and we find that the distribution is consistent with geography context information. As shown in Fig. 10 and Fig. 11, although the overall pattern is similar, we highlight some differences between the distributions of car users and public transportation users:

- Residents of upscale residential areas tend to travel by car. We identify several high-end residential areas (rectangle A in Fig. 10 and rectangle B in Fig. 11). The intensity of car users in these areas is significantly higher than that of public transportation users.
- The students accommodated in school tend to travel by public transportation. Rectangle B in Fig. 10 and Rectangle A in Fig. 11 framed the locations of multiple colleges and universities. In China, almost all college students live in school and do not have cars. They mainly take public transportation to go out.

In a nutshell, the accordance with census data and geography context suggests that CellTrans remains effective at urban scale.

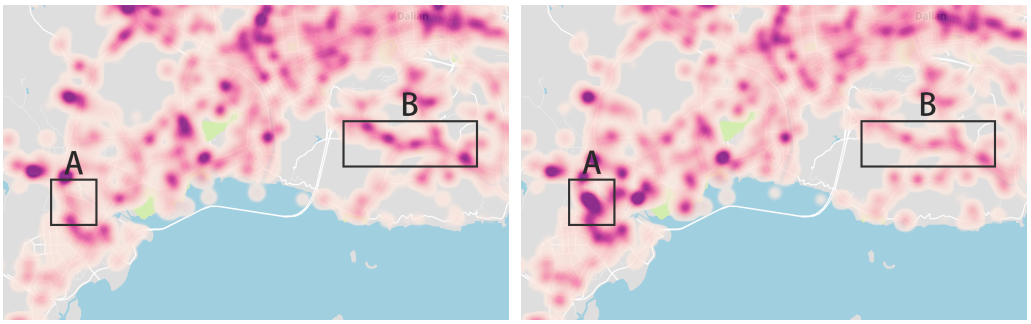
6.3 Scenario 2: Without Label Data

In Sec. 5.2, we present two methods to infer users' main transportation modes when there are no labeled users in the studied city: a clustering based method and transferred models from other cities. In this section, we will



(a) Car Users' Homes in Shenyang (b) Bus Users' Homes in Shenyang

Fig. 10. Heatmaps of the distribution of (a) car users' homes and (b) public transportation users' homes in Shenyang. Area A is one of Shenyang's upscale residential areas, where there are many car users as shown in (a). In comparison, the intensity for public transportation users is much lower. B is the location of a college, where the main residents are college students. In this area, there are much more public transportation users than car users.



(a) Car Users' Homes in Dalian (b) Bus Users' Homes in Dalian

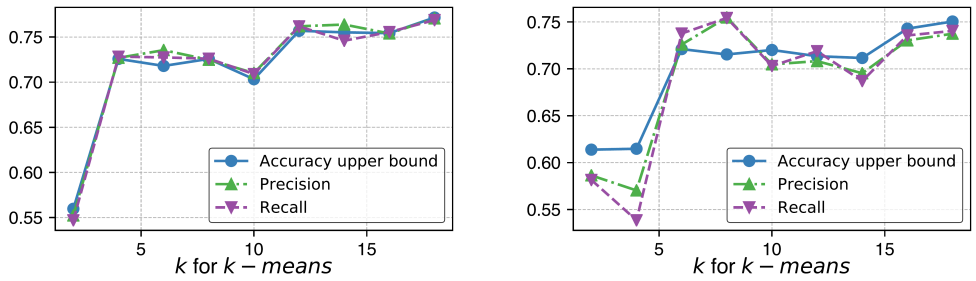
Fig. 11. Heatmaps of the distribution of (a) car users' homes and (b) public transportation users' homes in Dalian. There are three universities located in area A, where the main transportation modes of residents are public transportation. There are many high-end houses and apartments along the road covered by rectangle B. For the residents in area B, driving car is the more common transportation mode than public transportation.

evaluate their performance. First, we discuss the parameter selection for *k-means* in Sec. 6.3.1. Then we compare the performance of our methods with previous methods in Sec. 6.3.2.

6.3.1 Parameter Selection for k-means. We assess the accuracy of the clustering-based method utilizing the labeled users as discussed in Sec. 5.2. To set the parameter k (the number of clusters in *k-means*) properly, we observe how the upper bound of accuracy and the corresponding precision and recall vary with k. As shown in Fig. 12, the upper bound of accuracy is low when k is very small, so are the corresponding precision and recall. When k becomes larger, the three metrics all become higher, but the needed human efforts to infer the group centers' modes also grow. Based on Fig. 12, we set $k = 4$ for Shenyang and $k = 6$ for Dalian, because the accuracy gain is minor when k is larger than these values.

Table 9. Methods to infer main transportation mode.

Method	Description
CellTrans: *k-means*	Cluster users based on mobility features Sec. 4.2
CellTrans: Transferred model (Transfer)	Trained on labeled users from other cities
Aggregation: bus station matching (Station)	Trip information and bus station locations [19]
Aggregation: traveling time clustering (Cluster)	Trip information [35]
Route matching (Route)	Trips and navigation routes [27]



(a) Accuracy with k for Shenyang

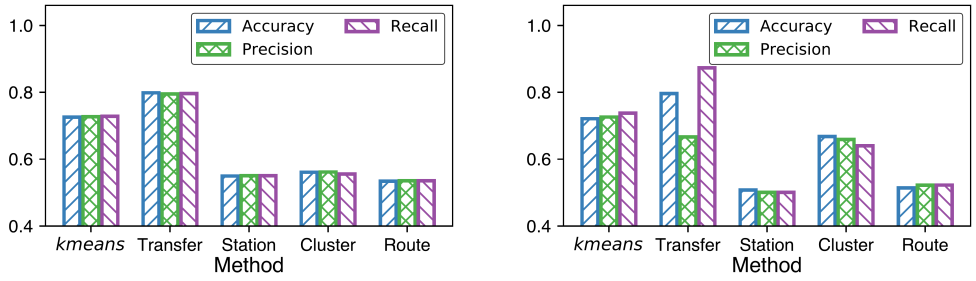
(b) Accuracy with k for Dalian

Fig. 12. X-axis is the number of clusters and y-axis is the corresponding accuracy, precision and recall.

6.3.2 Comparison with Previous Methods. If we want to survey users' main modes in a city where there are no labeled users, we present two options in Sec. 5.2: a clustering based method and transferred models from other cities. In this section, we compare their performance with previous methods. The comparing methods are listed in Table 9 and the results are shown in Fig. 13. For Shenyang, the model *Transfer* means the SVM model trained on Dalian users and for Dalian, it is the SVM model trained on Shenyang users. The kernel we use is still *radial basis function*. The three previous methods, *Station*, *Cluster* and *Route* are introduced before in Sec. 6.1.

The accuracy of our methods is higher than that of previous methods in both cities. The transferred model achieves the best results among all the methods. This suggests that the relationship between users' main transportation modes and their mobility features is consistent in different cities. *Station* and *Cluster* do not depend on the detailed velocity or acceleration information, which is almost impossible to extract from cellular data. They only consider simple trip features like the origin, destination and traveling time. However, car trips and bus trips are very similar concerning these features, which hold back these two methods to achieve high accuracy.

Route measures the distance between users' records and the navigation transit/driving routes retrieved from map service API (Amap [2] in our implementation). The reasons why it is not performing well are: (i) In a city with developed public transportation system, the navigation routes of buses and metros are very similar to the routes of driving car, especially for short distances. (ii) The irregular sampling of cellular data causes sparsity. The intervals between consecutive records can be as long as several hours. Cellular data can not delineate the detailed shape of trajectories due to these long blank intervals. (iii) The location error of cellular data can be several hundreds of meters. With such large error, cellular data are unable to capture the small difference between public transportation routes and driving car routes.



(a) Shenyang (b) Dalian

Fig. 13. The performance of different unsupervised methods. The methods are summarized in Table 9. (a) shows the results of Shenyang and (b) shows the results of Dalian.

In comparison, our methods consider the overall mobility features extracted from cellular data in a long period. Although the information provided by cellular data is coarse-grained and sparse, the accumulation over time can make up for cellular data's disadvantages and offer the chance to infer users' main transportation modes.

6.4 Feature Analysis

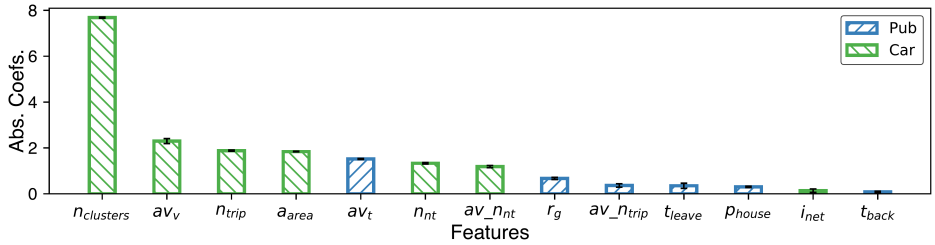
In this experiment, we analyze the importance and distribution of the features extracted in Sec. 4.2.

6.4.1 Rank Feature Importance. To find out the importance of features, we utilize the SVM with linear kernel, whose performance (81% accuracy for Shenyang and 84% for Dalian) is close to the SVM with radial basis function kernel. After fitting to data, the coefficients in linear SVM can reflect the importance of the corresponding features, *i.e.*, how much the feature contributes to separate the users into two classes [39].

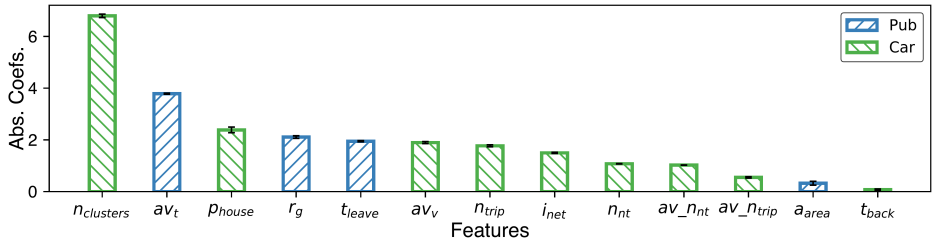
In Fig. 14, the absolute values of the coefficients are shown in descending order. The meanings of each feature can be found in Table 4. The larger the absolute coefficients, the more important the corresponding features are to infer main modes. The results show that: (1) The number of distinct stay clusters, average speed and time of trips, and the number of trips are important to separate users of different modes in both cities. (2) The convex hull area of stays is important in Shenyang, but less important in Dalian. (3) The house price, radius of gyration, the time to leave home and network access intensity are important in Dalian, but less important in Shenyang. (4) The time to come back home, the average number of trips per day are not important in both cities.

6.4.2 Feature Value Distribution. We analyze the distribution of the mobility features among users with different main modes. Some representative examples are shown in Fig. 15

- Some features have obviously different distribution between two modes, which depicts their discriminative ability. For example, $n_{cluster}$ is the most discriminative feature in both cities. As shown in Fig. 15a, it has different distribution between public transportation users and private car users. Private car users tend to visit more places than public transportation users.
- Some features have slightly different distribution and their differentiating ability is not as strong as the former ones. For example, Fig. 15c shows the distribution of av_t. The separation of two classes on av_t is not as obvious as $n_{cluster}$. This is consistent with the importance analysis in Sec. 6.4.1.
- Some features have similar distribution between different modes, but they still contain discriminating information based on the analysis in Sec. 6.4.1. For example, r_g is an important feature in Dalian, although



(a) Shenyang



(b) Dalian

Fig. 14. The absolute coefficients corresponding to features, reflecting each feature's importance. The error bar is the variance over 5-fold cross validation. The color represents the sign of the coefficients. Blue means that the larger the feature, the higher the probability that this user prefers public transportation. Green means the opposite.

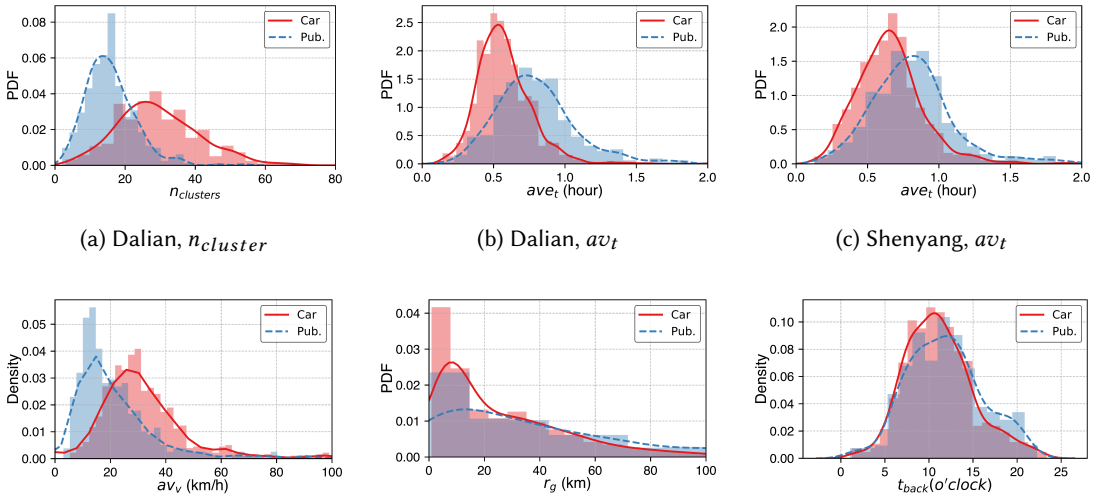
as shown in Fig. 15e, the distribution is very similar for two modes. The reason may be that although these features can not separate different modes individually, they are capable to do so combined with other features. That is to say, their distribution conditioned on other mobility features is more discriminative than their overall distribution.

- Lastly, some features contribute little to the differentiation of modes and naturally, they have very similar distribution. For example, t_{back} is the least important feature in both cities (Fig. 14). Its distribution is indistinguishable as shown in Fig. 15f.

6.5 Vary Temporal and Spatial Granularity

In the following experiments, we change the temporal and spatial granularity of the MFR datasets artificially and observe how the accuracy varies. We use the CellTrans with SVM and 5-fold cross-validation:

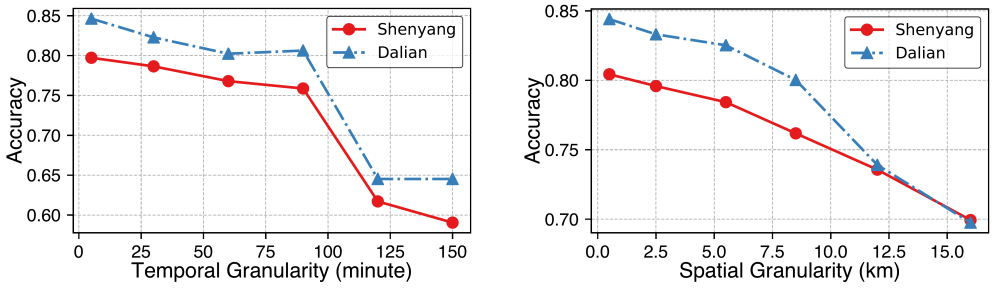
- **Temporal Granularity:** To change the data's temporal granularity to a specific value g_t, we examine each record $< t_i, p_i >$ of a user in chronological order (t_i is the time and p_i is the location of the record). If $t_i - t_{i-1} < g_t$, then record $< t_i, p_i >$ is discarded, otherwise we keep it and examine the next record. The results are shown in Fig. 16a. The accuracy encounters a significant drop near 100 minutes. We think the reason is that temporal granularity coarser than 100 minutes could entirely miss some traveling activities.
- **Spatial Granularity:** To change the data's spatial granularity to a specific value g_s, we examine each record $< t_i, p_i >$ of a user in chronological order. If $Dist(p_i, p_{i-1}) < g_s$, then we set the location of the i^{th} record to be the same as the $(i-1)^{th}$ record: $p_i \leftarrow p_{i-1}$. Otherwise we change nothing and examine the next record. The results are shown in Fig. 16b. Compared with temporal granularity, the accuracy of CellTrans decreases more smoothly when the spatial granularity becomes coarser.



(a) Dalian, $n_{cluster}$ (b) Dalian, av_t (c) Shenyang, av_t

(d) Dalian, av_v (e) Dalian, r_g (f) Dalian, t_{back}

Fig. 15. Feature distribution of car users and public transportation users.



(a) Temporal Granularity (b) Spatial Granularity

Fig. 16. Accuracy under different (a) temporal granularity and (b) spatial granularity. We manually change the temporal and spatial granularity of the data, and observe how the accuracy of CellTrans varies.

With these experiments we explore the lower bound of the granularity that CellTrans needs. The results suggest that CellTrans can work with data whose spatiotemporal granularity is even coarser than cellular data like MFR.

7 DISCUSSION

7.1 Lessons Learned

Based on the results of CellTrans, we learned several valuable lessons:

- The event-driven property of cellular data causes the irregular temporal sampling, which is a drawback for the study of human mobility. From another point of view, this can be an advantage that enable us to extract features about users' living patterns.
- Main transportation modes are not only related to users' trips, but also to their stays. Although inferring main transportation modes based on trips' modes does not work with coarse-grained cellular data, we can infer them based on users' general mobility features.
- The inference model for main transportation modes does not need a large number of training data, but it does need a long observation period (longer than 21 days). It is costly to collect groundtruth of many users, but it is easy for cellular dataset to cover a period longer than 3 weeks. So inferring main transportation modes with cellular data is a practical solution for urban-scale survey.
- The relationship between main transportation modes and mobility features is stable in different cities. A model trained in one city can be applied to other cities without much sacrifice of accuracy.

7.2 Limitation

- **Other transportation modes:** In this paper, we do not consider transportation modes like running, biking, taxi and so on. The reasons are: (i) We do not have the groundtruth for these transportation modes in our dataset. (ii) We focus on users' main transportation modes. People occasionally take taxies or run to go to some places, but few people take them as their main transportation modes. (iii) Driving car and taking public transportation are two most important transportation modes in urban lives. (iv) Cellular data localize users at cell granularity and the records are irregularly sampled. So for short-distance modes like walking, biking and running, we think it is almost impossible to differentiate them since their minor differences can not be reflected in cellular data.
- **Fine-grained public transportation:** We do not differentiate buses and subway/metro. They are included in *public transportation*. The main reason is that we do not have find-grained label data. However, we think CellTrans has the potential to deal with more fine-grained modes:
  - Current mobility features are possible to differentiate buses and subway/metro. For example, subway/metro usually have a higher speed than buses. *Average speed of trips* can capture this difference. The cellular signal is usually weak on subways and this can be characterized by *Network access intensity*.
  - Some new features can be added to enhance CellTrans to differentiate buses and subway/metro. For example, subway and metro have fixed stations. The closeness to these stations can be incorporated to identify subway/metro trips from others. Additionally, the high speed and underground may cause unique tower switch patterns.

7.3 Applications

- **For mobile phone service providers:** Mobility patterns and transportation modes affect the cell tower handover frequency, paging activities, path switching, etc. So inferring the transportation mode of cellular users is important for service providers to improve their services. Knowing users' main transportation modes can also help with user profiling: (i) Car ownership can reflect a user's economic status, which is one of the most critical factors in precision marketing. (ii) the transportation mode can help with understanding users' cellular traffic needs. When people are on buses, they usually use phones more often to kill time.
- **For governments:** Transportation mode information tells us the transportation needs of citizens, which is crucial to city planning and transportation management. CellTrans can also be integrated with the travel demmand forecasting models. Travel demand forecasting system considers multiple aspects about users' travel behaviors, including car ownership, trip frequency, destination choice, route choice and mode choice. CellTrans can support current travel demand forecasting in following aspects:

Table 10. Comparison with Related Work. *Level* considers whether the method focuses on trips' transportation mode or users' main transportation mode. *Car&Pub* means if the method differentiates between car and public transportation. *TW* represents for *Time Window*

	Level	Urban scale	#Labeled users	Data (#Users)	Car&Pub.
[5, 49, 51]	Trip	No	182	GPS (182)	Yes
[34]	Trip	No	125	GPS and user profiles	Yes
[4, 12, 14, 18, 20, 32]	TW	No	3-20	Sensor data (3-20)	Yes
[33]	TW	No	224	Sensor data (224)	Yes
[19]	Trip	Yes	0	Cellular and bus data (unknown)	Only Pub.
[35]	Trip	No	0	Cellular data (one million)	Yes
[27]	Trip	No	0	Cellular data (5405)	Yes
[41]	Trip	No	500	Cellular (500)	No
CellTrans	**Main**	**Yes**	**2,589**	**Cellular data (three million)**	**Yes**

- CellTrans can provide a priori probability distribution of users' mode preference. This distribution can be integrated into current mode predicting models.
- The results of CellTrans can be aggregated at regional level to provide the transport mode profile of the region. This profile can be used in inter-region traffic predicting.
- Car ownership is closely related to the user's main transport mode. If a user's main transport mode is private car, he/she is very likely to have a car.

7.4 Privacy Protection

The cellphone data are collected by cellular service providers, and we access the data as academic collaborators. By signing cellular service contracts, all the cellphone users consent that their metadata will be used to analyze how they use cellphone infrastructures for performance profiling, anomaly detection, better services, etc. As discussed in Sec. 7.3, CellTrans can help providers improve their services for cellphone users.

While the analysis of cellular data has great potential for social benefits, we must take active actions to protect users' privacy: (i) All users' IDs have been hashed into random yet unique identifiers by the cellular service providers' engineers who did not participate in this work. (ii) The original data were never moved out of our research facility, no data access to the personnel not related to this project. (iii) We only process the data fields that are useful in this project, and drop others for minimal exposure. (iv) The final results are the main transportation modes of users. The results do not contain the specific places that users visit. This paper is a part of an industry collaborative project within our university, the paper publication with the real-world data-driven experiments and results is allowed.

8 RELATED WORK

Our work is directly related to previous works that infer transportation modes from heterogeneous data sources and is inspired by the various applications of cellular data in urban sensing. So we reviewed the related works from the following two aspects:

8.1 Transportation Mode Inference

Understanding citizens' transportation mode preference is beneficial to a range of applications. Decades ago, questionnaires were the main means to gather information about users' transportation modes. Then the emergence

of GPS-equipped devices provides a convenient and detailed way to collect users' location information. Thus many researchers study how to infer the transportation modes of GPS trajectories. For example, some works extract velocity related features and feed them into supervised models [31, 49, 51], Wang et al. consider both mobility and socioeconomic features, and some works design deep learning models to study transportation modes: Dabiri and Heaslip devise a CNN model for mode inference and Song et al. utilize LSTM for the prediction and simulation of transportation modes. Besides GPS, other sensors like accelerators, gyroscopes and so on are also used to infer users' real-time transportation modes with different methods [4, 12, 14, 18, 20, 32, 33].

Nevertheless, these methods are limited to a small number of users because of the difficulty of data collection. In recent years, cellular networks provide new opportunities to collect large scale location information at low cost, and several works try to infer trips' transportation modes from cellular data [19, 27, 35, 41]. However, the low spatiotemporal granularity makes it failed to infer users' main modes by first analyzing users' trips' modes. In this paper, we design a new framework to survey users' main transportation modes at urban scale by utilizing the mobility features extracted from cellular data.

The comparison of related works is summarized in Table 10. We evaluate CellTrans with the largest dataset and most groundtruth users.

8.2 Urban Sensing with Cellular Networks

The high penetration and the ubiquitous coverage make cellular networks an ideal infrastructure for comprehensive and large scale urban sensing. Calabrese et al. give a thorough survey in this area. In recent years, with more and more accessible datasets, new methods and new applications have come forth continuously: Fang et al. model urban population with multiple cellular networks to compensate for the bias of a single network. Isaacman et al. identify the important places for users based on cellular data. There are also some works studying the fundamental laws of human mobility [11, 21, 29]. Besides human mobility, researchers also consider the variation of traffic volume at urban scale, *e.g.*, Wang et al. predict the traffic volume at tower level, Wang et al. study the urban mobile traffic patterns, and Ferrari et al. try to detect the special events in the city by analyzing cellular traffic. Additionally, mobile phones have been widely used in mobile crowd sensing [46, 47]. Our work profiles users' main transportation modes at urban scale, which extends the applications of cellular data in urban sensing.

9 CONCLUSION

In this paper, we present CellTrans, a novel cellular network based framework to survey users' main transportation modes (public transportation or private car) at urban scale. We devise techniques to extract mobility features from noisy MFRs that are pertinent to users' transportation modes. To make CellTrans more practical and generalizable, we consider different application scenarios and design solutions correspondingly. We evaluate the performance of CellTrans on two large-scale MFR datasets from two metropolises in the northeast of China, which cover 3 million users. We carry out comprehensive experiments to evaluate the performance of CellTrans quantitatively and qualitatively. The experimental results show that CellTrans can achieve high accuracy with or without labeled users in the studied cites, and CellTrans remains effective when applied at urban scale to the whole population.

ACKNOWLEDGMENTS

This work is supported in part by the National Key Research Plan under Grant No.: 2016YFC0700100, NSFC under Grant No.: 61832010, 61632008, 61672319 and 61872081, Microsoft Research Asia, and Tsinghua University Initiative Scientific Research Program.

REFERENCES

[1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C. González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58 (2015), 240 – 250. https://doi.org/10.1016/j.trc.2015.

02.018 Big Data in Transportation and Traffic Engineering.
[2] Amap. 2019. Amap Open Platform. Retrieved May 11, 2019 from https://lbs.amap.com/
[3] Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. 2014. Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Comput. Surv.* 47, 2, Article 25 (Nov. 2014), 20 pages. https://doi.org/10.1145/2655691
[4] Ke-Yu Chen, Rahul C. Shah, Jonathan Huang, and Lama Nachman. 2017. Mago: Mode of Transport Inference Using the Hall-Effect Magnetic Sensor and Accelerometer. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 8 (June 2017), 23 pages. https://doi.org/10.1145/3090054
[5] Sina Dabiri and Kevin Heaslip. 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation Research Part C: Emerging Technologies* 86 (2018), 360 – 371. https://doi.org/10.1016/j.trc.2017.11.021
[6] M. G. Demissie, S. Phithakkitnukoon, and L. Kattan. 2018. Trip Distribution Modeling Using Mobile Phone Data: Emphasis on Intra-Zonal Trips. *IEEE Transactions on Intelligent Transportation Systems* (2018), 1–13. https://doi.org/10.1109/TITS.2018.2868468
[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, 226–231. http://dl.acm.org/citation.cfm?id=3001460.3001507
[8] 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
[9] L. Ferrari, M. Mamei, and M. Colonna. 2012. People get together on special events: Discovering happenings in the city via cell network analysis. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. 223–228. https://doi.org/10.1109/PerComW.2012.6197484
[10] H. Gjoreski, M. Ciliberto, L. Wang, F. J. Ordonez Morales, S. Mekki, S. Valentin, and D. Roggen. 2018. The University of Sussex-Huawei Locomotion and Transportation Dataset for Multimodal Analytics With Mobile Devices. *IEEE Access* 6 (2018), 42592–42604. https://doi.org/10.1109/ACCESS.2018.2858933
[11] Marta C. González, César A. Hidalgo, and Albert-László Barabási. 2008. Understanding individual human mobility patterns. *Nature* 453 (2008), 779. https://doi.org/10.1038/nature06958
[12] Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2013. Accelerometer-based Transportation Mode Detection on Smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys '13)*. ACM, New York, NY, USA, Article 13, 14 pages. https://doi.org/10.1145/2517351.2517367
[13] 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 *Pervasive Computing*, Kent Lyons, Jeffrey Hightower, and Elaine M. Huang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 133–151.
[14] Jeya Vikranth Jeyakumar, Eun Sun Lee, Zhengxu Xia, Sandeep Singh Sandha, Nathan Tausik, and Mani Srivastava. 2018. Deep Convolutional Bidirectional LSTM Based Transportation Mode Recognition. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. ACM, New York, NY, USA, 1606–1615. https://doi.org/10.1145/3267305.3267529
[15] S. Jiang, J. Ferreira, and M. 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 (June 2017), 208–219. https://doi.org/10.1109/TBDATA.2016.2631141
[16] Yu Jin, Nick Duffield, Alexandre Gerber, Patrick Haffner, Wen-Ling Hsu, Guy Jacobson, Subhabrata Sen, Shobha Venkataraman, and Zhi-Li Zhang. 2012. Characterizing Data Usage Patterns in a Large Cellular Network. In *Proceedings of the 2012 ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design (CellNet '12)*. ACM, New York, NY, USA, 7–12. https://doi.org/10.1145/2342468.2342471
[17] Kevin S. Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. 2014. Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data. *PLOS ONE* 9, 6 (06 2014), 1–15. https://doi.org/10.1371/journal.pone.0096180
[18] G. Lan, W. Xu, S. Khalifa, M. Hassan, and W. Hu. 2016. Transportation mode detection using kinetic energy harvesting wearables. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 1–4. https://doi.org/10.1109/PERCOMW.2016.7457048
[19] Guanyao Li, Chun-Jie Chen, Sheng-Yun Huang, Ai-Jou Chou, Xiaochuan Gou, Wen-Chih Peng, and Chih-Wei Yi. 2017. Public Transportation Mode Detection from Cellular Data. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 2499–2502. https://doi.org/10.1145/3132847.3133173
[20] Jonathan Liono, Zahraa S. Abdallah, A. K. Qin, and Flora D. Salim. 2018. Inferring Transportation Mode and Human Activity from Mobile Sensing in Daily Life. In *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '18)*. ACM, New York, NY, USA, 342–351. https://doi.org/10.1145/3286978.3287006
[21] Tongtong Liu, Zheng Yang, Yi Zhao, Chenshu Wu, Zimu Zhou, and Yunhao Liu. 2018. Temporal understanding of human mobility: A multi-time scale analysis. *PLOS ONE* 13, 11 (2018), e0207697.
[22] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica. 2016. Large-Scale Mobile Traffic Analysis: A Survey. *IEEE Communications Surveys Tutorials* 18, 1 (Firstquarter 2016), 124–161. https://doi.org/10.1109/COMST.2015.2491361

[23] Dalian Municipal Bureau of Statistics. 2017. 2016 Statistical Communique of National Economic and Social Development in Dalian. Retrieved January 26, 2019 from http://www.stats.dl.gov.cn/index.php?m=content&c=index&a=show&catid=52&id=12000

[24] Shenyang Municipal Bureau of Statistics. 2017. 2016 Statistical Communique of National Economic and Social Development in Shenyang. Retrieved January 26, 2019 from http://www.shenyang.gov.cn/zwgk/system/2017/09/14/010193052.shtml

[25] Dalian Xinshang Newspaper Office. 2017. There needs more than 1.04 million parking berths for Dalian's 1.52 million vehicles. Retrieved January 26, 2019 from http://dl.sina.com.cn/news/m/2017-01-10/detail-ifxzkfuh6518304.shtml

[26] Santi Phithakkitnukoon, Zbigniew Smoreda, and Patrick Olivier. 2012. Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. *PLOS ONE* 7, 6 (06 2012), 1–9. https://doi.org/10.1371/journal.pone.0039253

[27] Santi Phithakkitnukoon, Titipat Sukhvibul, Merkebe Demissie, Zbigniew Smoreda, Juggapong Natwichai, and Carlos Bento. 2017. Inferring social influence in transport mode choice using mobile phone data. *EPJ Data Science* 6, 1 (14 Jun 2017), 11. https://doi.org/10.1140/epjds/s13688-017-0108-6

[28] Ling Qi, Yuanyuan Qiao, Fehmi Ben Abdesslem, Zhanyu Ma, and Jie Yang. 2016. Oscillation Resolution for Massive Cell Phone Traffic Data. In *Proceedings of the First Workshop on Mobile Data (MobiData '16)*. ACM, New York, NY, USA, 25–30. https://doi.org/10.1145/2935755.2935759

[29] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. Modelling the scaling properties of human mobility. *Nature Physics* 6 (2010), 818. https://doi.org/10.1038/nphys1760

[30] 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 (IJCAI'16)*. AAAI Press, 2618–2624. http://dl.acm.org/citation.cfm?id=3060832.3060987

[31] Leon Stenneth, Ouri Wolfson, Philip S. Yu, and Bo Xu. 2011. Transportation Mode Detection Using Mobile Phones and GIS Information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '11)*. ACM, New York, NY, USA, 54–63. https://doi.org/10.1145/2093973.2093982

[32] Ritiz Tambi, Paul Li, and Jun Yang. 2018. An Efficient CNN Model for Transportation Mode Sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems (SenSys '18)*. ACM, New York, NY, USA, 315–316. https://doi.org/10.1145/3274783.3275160

[33] Toan H. Vu, Le Dung, and Jia-Ching Wang. 2016. Transportation Mode Detection on Mobile Devices Using Recurrent Nets. In *Proceedings of the 24th ACM International Conference on Multimedia (MM '16)*. ACM, New York, NY, USA, 392–396. https://doi.org/10.1145/2964284.2967249

[34] B. Wang, L. Gao, and Z. Juan. 2018. Travel Mode Detection Using GPS Data and Socioeconomic Attributes Based on a Random Forest Classifier. *IEEE Transactions on Intelligent Transportation Systems* 19, 5 (May 2018), 1547–1558. https://doi.org/10.1109/TITS.2017.2723523

[35] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti. 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *13th International IEEE Conference on Intelligent Transportation Systems*. 318–323. https://doi.org/10.1109/ITSC.2010.5625188

[36] Huandong Wang, Fengli Xu, Yong Li, Pengyu Zhang, and Depeng Jin. 2015. Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. ACM, New York, NY, USA, 225–238. https://doi.org/10.1145/2815675.2815680

[37] Lin Wang, Hristijan Gjoreskia, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2018. Summary of the Sussex-Huawei Locomotion-Transportation Recognition Challenge. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. ACM, New York, NY, USA, 1521–1530. https://doi.org/10.1145/3267305.3267519

[38] X. Wang, Z. Zhou, Z. Yang, Y. Liu, and C. Peng. 2017. Spatio-temporal analysis and prediction of cellular traffic in metropolis. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*. 1–10. https://doi.org/10.1109/ICNP.2017.8117559

[39] Wikipedia. [n. d.]. Support-vector Machine. https://en.wikipedia.org/wiki/Support-vector_machine#Linear_SVM

[40] Fang Tian Xia. 2019. Fang Tian Xia. Retrieved January 26, 2019 from https://sy.fang.com/

[41] Dafeng Xu, Guojie Song, Peng Gao, Rongzeng Cao, Xinwei Nie, and Kunqing Xie. 2011. Transportation Modes Identification from Mobile Phone Data Using Probabilistic Models. In *Advanced Data Mining and Applications*, Jie Tang, Irwin King, Ling Chen, and Jianyong Wang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 359–371.

[42] 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 (UbiComp '16)*. ACM, New York, NY, USA, 1064–1075. https://doi.org/10.1145/2971648.2971673

[43] P. Yang, T. Zhu, X. Wan, and X. Wang. 2014. Identifying Significant Places Using Multi-day Call Detail Records. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*. 360–366. https://doi.org/10.1109/ICTAI.2014.61

[44] Su Yang, Minjie Wang, Wenshan Wang, Yi Sun, Jun Gao, Weishan Zhang, and Jiulong Zhang. 2017. Predicting Commercial Activeness over Urban Big Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 119 (Sept. 2017), 20 pages. https://doi.org/10.1145/3130983

[45] Paul A Zandbergen. 2009. Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. *Transactions in GIS* 13, s1 (2009), 5–25. https://doi.org/10.1111/j.1467-9671.2009.01152.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9671.2009.01152.x

[46] X. Zhang, Z. Yang, Y. Liu, and S. Tang. 2019. On Reliable Task Assignment for Spatial Crowdsourcing. *IEEE Transactions on Emerging Topics in Computing* 7, 1 (Jan 2019), 174–186. https://doi.org/10.1109/TETC.2016.2614383

[47] Xinglin Zhang, Zheng Yang, Wei Sun, Yunhao Liu, Shaohua Tang, Kai Xing, and Xufei Mao. 2016. Incentives for mobile crowd sensing: A survey. *IEEE Communications Surveys & Tutorials* 18, 1 (2016), 54–67.

[48] Yi Zhao, Zimu Zhou, Xu Wang, Tongtong Liu, Yunhao Liu, and Zheng Yang. 2019. CellTradeMap: Delineating Trade Areas for Urban Commercial Districts with Cellular Networks. In *IEEE International Conference on Computer Communications (INFOCOM 2019)*.

[49] 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, Article 1 (Jan. 2010), 36 pages. https://doi.org/10.1145/1658373.1658374

[50] 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 (UbiComp '08)*. ACM, New York, NY, USA, 312–321. https://doi.org/10.1145/1409635.1409677

[51] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning Transportation Mode from Raw Gps Data for Geographic Applications on the Web. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 247–256. https://doi.org/10.1145/1367497.1367532

[52] Yu Zheng, Xing Xie, and Wei-Ying Ma. 2010. GeoLife: A Collaborative Social Networking Service among User, location and trajectory. *IEEE Data(base) Engineering Bulletin* (June 2010). https://www.microsoft.com/en-us/research/publication/geolife-a-collaborative-social-networking-service-among-user-location-and-trajectory/

[53] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 791–800. https://doi.org/10.1145/1526709.1526816