9-2020

Urban scale trade area characterization for commercial districts with cellular footprints

Yi ZHAO

Zimu ZHOU
Singapore Management University, zimuzhou@smu.edu.sg

Xu WANG

Tongtong LIU

Zheng YANG

Urban Scale Trade Area Characterization for Commercial Districts with Cellular Footprints*

YI ZHAO, School of Software and BNRist, Tsinghua University, China
ZIMU ZHOU, School of Information Systems, Singapore Management University, Singapore
XU WANG, School of Software and BNRist, Tsinghua University, China
TONGTONG LIU, School of Software and BNRist, Tsinghua University, China
ZHENG YANG†, School of Software and BNRist, Tsinghua University, China

Understanding customer mobility patterns to commercial districts is crucial for urban planning, facility management, and business strategies. Trade areas are a widely applied measure to quantify where the visitors are from. Traditional trade area analysis is limited to small-scale or store-level studies because information such as visits to competitor commercial entities and place of residence is collected by labour-intensive questionnaires or heavily biased location-based social media data. In this paper, we propose CellTradeMap, a novel district-level trade area analysis framework using mobile flow records (MFRs), a type of fine-grained cellular network data. We show that compared to traditional cellular data and social network check-in data, MFRs can model customer mobility patterns comprehensively at urban scale. CellTradeMap extracts robust location information from the irregularly sampled, noisy MFRs, adapts the generic trade area analysis framework to incorporate cellular data, and enhances the original trade area model with cellular-based features. We evaluate CellTradeMap on two large-scale cellular network datasets covering 3.5 million and 1.8 million mobile phone users in two metropolis in China respectively. Experimental results show that the trade areas extracted by CellTradeMap are aligned with domain knowledge and CellTradeMap can model trade areas with a high predictive accuracy.

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

Additional Key Words and Phrases: cellular networks, crowdsensing, trade area analysis, human mobility

ACM Reference Format:
Yi Zhao, Zimu Zhou, Xu Wang, Tongtong Liu, and Zheng Yang. 2018. Urban Scale Trade Area Characterization for Commercial Districts with Cellular Footprints. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/1122445.1122456

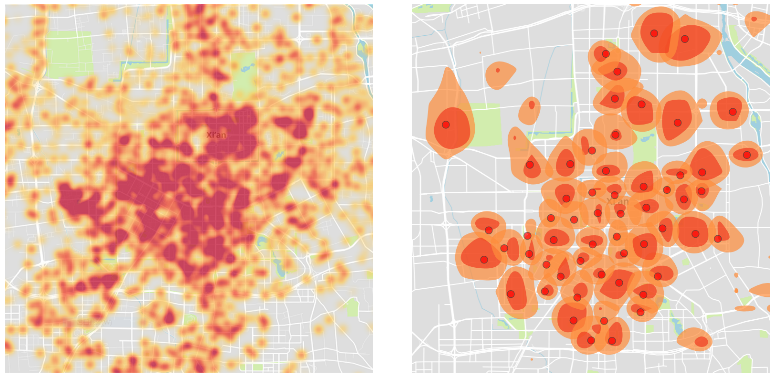
1 INTRODUCTION

The ubiquity of mobile devices and the development of cellular networks have generated unprecedented telecommunication big data. There have been more mobile devices than humans

*A preliminary version of this article appeared in International Conference on Computer Communications (IEEE INFOCOM 2019)
†Corresponding author: Zheng Yang.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456



(a) (b)

Fig. 1. (a) Spatial distributions of MFRs in one hour. (b) Trade areas of all commercial districts in a city. Red circles are commercial districts. The contour maps around circles are the corresponding trade areas.

worldwide [2]. These devices access cellular networks for various applications, such as news browsing, instant messages, mobile videos, mobile games, *etc.* It is predicted that the annual mobile Internet traffic will exceed half a ZB (10^{21}Bytes) by 2021 [5].

The tremendous amounts of cellular network records contain precious business values. Cellular data have long served as approximated locations of mobile users at the granularity of cell towers [27, 37]. Over the past decade, researchers have exploited cellular data to mine customer mobility behaviour for various business strategies and applications such as mobile advertising [9], optimal store location planning [17] and commercial activeness prediction [33].

One expressive, widely adopted approach to characterize customer mobility pattern is *trade areas*. A trade area is "a geographically delineated region containing potential customers", which quantifies the distributions of visitors to a store or a commercial district [14]. In other words, the trade area of a store or a commercial district depicts the origins (*i.e.*, home locations) of visitors and the corresponding visit probabilities. Understanding where the visitors come from and their choices of competitive stores or commercial districts is vital to optimize market management and strategies.

Despite its importance, trade area analysis has long been considered expensive and time-consuming. The major burden is the efforts to estimate the number of visitation to a store or commercial district and all of its competitors, as well as to collect home information of the visitors. Traditionally, such information is manually collected from questionnaires and surveys [31]. Researchers interview the residents in an area to know how often people visit commercial areas and which commercial districts they visit. Such methods are laborious and limited in small scale.

Other studies [17, 23, 31] utilize location data from social media as alternative method for trade area analysis. Check-in data from social media prove effective due to their ease to be collected at large scale [23]. However, infering place of home and collecting comprehensive visitation information of competitor businesses are very difficult based on the biased and limited check-in data [18] Furthermore, it is difficult to aggregate the trade area of stores to obtain the trade area of commercial districts without bias. Call Detail Records(CDR) can also provide location information, but only when users make or receive phone calls. This make CDR data very sparse and unable to support detailed inspection such as trade area analysis.

To fill the void of cost-effective, urban-scale, comprehensive trade area analysis for commercial districts, we explore mobile flow records (MFR), a fine-grained cellular network data source that has recently attract much research attention [20]. MFRs are system logs of cellular network that describe the Internet access behaviour of phone users. The wide spatial coverage (*e.g.* 3.5 million mobile phone users in a metropolis) and high time resolution (*e.g.* 4-minute sampling rate) make them suited for comprehensive district-level trade area analysis. In comparison, effective check-ins may be sparse and contain data for limited numbers and types of stores (*e.g.* 4 stores in New York with 0.1 check-in per user per day. [23]).

We propose CellTradeMap, an MFR-based framework to delineate and model trade areas for commercial districts. We base our design upon large-scale MFR datasets covering millions of anonymous mobile phone users in two metropolises of China, which makes urban-scale analysis possible. Through measurement studies, we investigate the irregular sampling and frequent base station switch problems of the MFR data. To tackle these challenges, we design a novel and practical pipeline to extract robust location information in the form of stay points from raw MFRs. We also adapt the generic trade area analysis framework [25] to incorporate this cellular data, extend the scope of trade area analysis to various attractiveness metrics, and improve the accuracy of the widely adopted trade area model [14] by adding MFR-based metrics and $L^1 - norm$. Fig. 1a shows the spatial distribution of our MFR dataset within an hour, and Fig. 1b illustrates the trade areas of all the commercial districts in the city derived from CellTradeMap in the form of contour maps.

We summarize the main contributions of this work below.

- To the best of our knowledge, this is the first work that utilizes flow-level data of cellular networks to profile and model trade areas for commercial districts. It offers a new cost-effective data collection methodology for urban-scale district-level trade area analysis.
- We design practical processing techniques to extract stay points and home locations of users from raw MFR data. Our solution serves as a generic pipeline to robustly derive location information from flow-level cellular data for mobility-related studies.
- We adapt the general trade area analysis framework to incorporate MFR and conduct urban-scale analysis on an MFR dataset. Experiments show that CellTradeMap profiles trade areas that are explainable by prior knowledge, reveal the important metrics for commercial attractiveness, and improves the predictive accuracy of the conventional trade area model with the help of $L^1 - norm$ and MFR-based metrics.

A preliminary version of CellTradeMap has been presented in [38]. We extend it in the following aspects:

- We evaluate the performance of CellTradeMap on a new dataset (Sec. 7), which covers 1.8 million mobile phone users in 48 days. The results further show that CellTradeMap can delineate and model trade areas effectively.
- We compare the results of two different cities and find that some attractive metrics are consistent in different cities and some differ. (Sec. 7.3)
- We compare MFR with Call Detail Records (CDR) (Sec. 3.1) and check-in data (Sec. 7.4), which are widely used in previous work [23, 28, 31]. The results show that MFRs are more suitable to analyze customers' behaviors.

In the rest of this paper, we review related work in Sec. 2, introduce our dataset and CellTradeMap framework in Sec. 3, present the details of the three modules of CellTradeMap in Sec. 4, Sec. 5 and Sec. 6, and evaluate its performance in Sec. 7. Finally we conclude this paper in Sec. 8.

2 RELATED WORK

Our work is inspired by the emerging trend on urban sensing with cellular networks, with a focus on trade area analysis. We review the most relevant studies below.

2.1 Urban Sensing with Cellular Networks

High user penetration and large spatial coverage make cellular networks an ideal data source for large-scale and comprehensive urban sensing [4, 20, 35, 36]. Different types of cellular data for various applications has been exploited in previous studies.

Aggregated traffic data have been studied to monitor and manage urban cellular traffic. Ferrari *et al.* [12] partition the urban area into grids and agglomerate cellular usage data in each grid to detect events in city. Wang *et al.* [30] study at cellular tower level to predict future traffic in the city.

Call detail record (CDR) is another type of cellular data, which records a time stamp and the connected tower ID when a phone call is made. It contains information about users' location and have been used to study the fundamental laws of human mobility [13, 27, 28]. [34] combile CDR data and public transit data to infer human mobility patterns more accurately.

MFR are sampled whenever mobile phones accesses the cellular network, which contains much more detailed information than CDR. Several previous works use MFR for fine-grained traffic characterization [29] and mobility modeling [19, 37]. Our work is the first to devise techniques for MFR to infer the locations of residence and visits to commercial districts for trade area analysis.

2.2 Trade Area Analysis

Trade area analysis studies questions such as "how long distance did people travel" and "what factors attract customers" to a store or a commercial district. Understading these questions can help with city planning and market management [24]. To do trade area analysis, researchers need to estimate the number of visitation to stores or commercial districts. Traditionally, these information is collected by surveys[10, 21].

User check-ins on social networks emerge as a low-cost alternative to estimate the number of visitation [17, 23, 31]. Wang *et al.* [23] characterize where the customers of four popular stores come from exploiting check-in data of the four stores in New York City. Wang *et al.* [31] highlight the effects of different customer sample sets on trade area analysis by investigating check-in data of five major commercial districts in Beijing, China. However, check-in data suffers from the sparsity and bias problems [23], making them unfit for comprehensive trade area analysis at district level. This also limits their ability to quantify the metrics' impact on trade areas.

We conduct trade area analysis with MFRs, which have wider spatial coverage and finer temporal resolution than check-ins, and design processing techniques dedicated to extract robust location information from MFRs.

3 OVERVIEW

This section presents our mobile flow record dataset and the overall framework of CellTradeMap.

3.1 Mobile Flow Record Dataset

Mobile flow records (MFRs) are fine-grained logs of cellular networks. Each MFR consists of a user ID, a time stamp, the base station ID, the application sending this packet, the host and the Uniform Resource Identifier (URI) of the request, as well as other flow information like upload/download bytes (o2r/r2o in Table 1). We use two MFR datasets (*D1* and *D2*) to evaluate the performance of CellTradeMap. They are from two different cities in China and both cover a broad area and a large population. Their statistics are summarized in Table 2.

Table 1. Example of mobile flow record.

User	Time	Station	Host	URI	o2r Bytes	r2o Bytes	...
$user_1$	t_1	s_1	www.example.com	/index/...	614	418	...
...	...	...	...	...	...	...	...

Table 2. Dataset Statistics

	D1	*D2*
# Records	1.7×10^{10}	8×10^9
# Cell towers	2.1×10^4	1.2×10^4
# Covered users	3.5×10^6	1.8×10^6
Covered area	1.1×10^4 km^2	1.3×10^4 km^2
Covered period	June 6 - 18, 2016	Dec. 19, 2016 - Feb. 4, 2017

The MFR datasets cover a wide spatial range and have a high time resolution. In comparison, the check-in dataset used in [23] only contain data for 4 stores in New York City (around 100 check-ins for each store). The average number of records per user per day of our MFR data is 694 and the average interval is shorter than 4 minutes. In contrast, the average number of check-ins per user per day in [23] is only 0.1.

Compared to CDR, which is commonly used in previous works (Sec. 2.1), MFR has much finer temporal granularity and more records. We compare MFR and CDR in Fig. 2:

- Fig. 2a compares the temporal granularity of MFR and CDR by the CDF of the inter-record intervals. Most intervals of MFR are shorter than several minutes, while the inter-record interval of CDR can be as long as several hours. As shown in Fig. 2a, nearly all consecutive MFRs are within 10^3 seconds (about 17 minutes). For CDR, there are over one forth inter-record intervals that are longer than 10^4 seconds (about 3 hours).
- Fig. 2b shows the CDF of the distance between consecutive distinct location records. The spatial granularity of CDR is close to that of MFR because it is mainly decided by the density of base stations. The slightly coarser spatial granularity of CDR may be due to the long inter-record intervals.
- Fig. 2c is the CDF of the number of daily records per user. 95% users have less than 1 call detail record per day, while most users have hundreds of mobile flow records in one day on average. This limits the capacity of CDR to characterize users' daily activities like visiting commercial districts.

The high user penetration and fine temporal granularity make MFR ideal to survey users' visits to commercial districts and their places of residence, which are the basis for the trade area analysis.

Together with other data sources such as Points of Interest (POIs), MFRs hold potential to comprehensively analyze the trade areas for commercial districts in the entire city.

3.2 CellTradeMap Framework

CellTradeMap is a new pipeline to characterize and predict the trade areas for commercial districts with MFRs. It consists of three major functional modules (see Fig. 3).

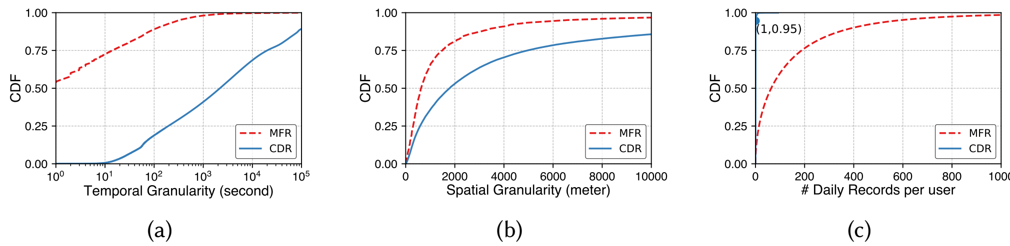


Fig. 2. Compare CDR and MFR on (a) temporal granularity, (b) spatial granularity, and (c) the number of daily records per user. MFR has much finer temporal granularity, slightly finer spatial granularity and more records per user per day. The CDF of MFR is based on *D1* and the CDF of CDR is based on a CDR dataset that covers 4.4×10^5 users in 27 days.

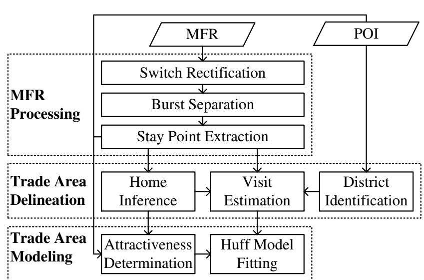
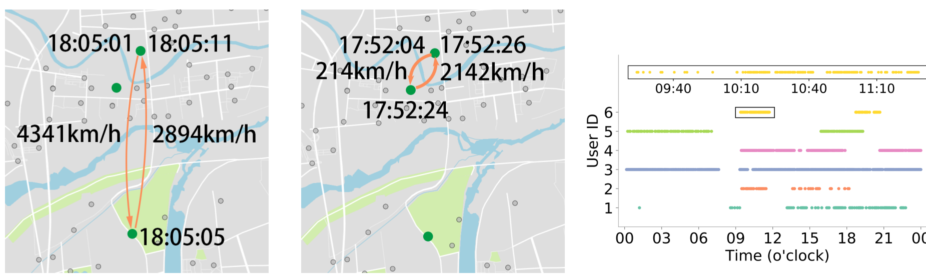


Fig. 3. Overview of CellTradeMap.

- **MFR Processing.** In this module, we extract *stay points and durations* from mobile phone users' raw MFRs. Recent proposals exploit check-ins from social media to count visitations [17], but such data is prone to sparsity and bias [23]. Techniques to extract stay points from GPS traces [39, 40] cannot be applied to MFRs because of MFRs' unique characteristics (Sec. 4.1). We design novel processing pipeline including switch rectification, burst separation and stay points extraction, to robustly extract location and visitation information from MFRs in Sec. 4.
- **Trade Area Delineation.** This module visualizes the trade areas *e.g.*, with contour maps of visit probabilities (see Fig. 1b). We harness POI clustering to identify commercial districts, infer home locations of visitors based on spatiotemporal patterns of MFRs, and estimate visit probabilities to commercial districts (Sec. 5). We also explain the different patterns of trade areas (Sec. 7.2.2).
- **Trade Area Modeling.** This module associates contexts such as the attractiveness of a commercial district to its visit probability. The Huff gravity model [14] is widely used to predict the trade area of commercial districts. However, there is no consensus on a unified definition of the attractiveness. We extract new metrics from MFRs and POIs to quantify the attractiveness, evaluate each metric's contribution to attractiveness and improve the accuracy of the original Huff model (Sec. 6).

In the next three sections, we detail each of the three functional modules in sequel.



(a) (b) (c)

Fig. 4. (a)(b): Switches to (a) a remote base station and (b) a nearby base station. The green (bigger) dots are the user's connected stations at the corresponding time, and the grey (smaller) dots are all stations nearby. (c): Bursty sampling of MFRs. Points on each horizontal line represent the occurrences of one user's MFRs. One of User 6's bursts is zoomed in at the top.

4 MOBILE FLOW RECORD PROCESSING

This section presents the pipeline to robustly extract stay points of mobile phone users from MFRs.

4.1 Challenges

4.1.1 Frequent Base Station Switches. MFRs are expected to approximate users' location by the connected base station's location. In practice, the phone is not always connected to the nearest base station because of the overlap of base stations' service areas [22]. Sometimes a phone may suddenly connect to a remote base station, exchange several packets and switch back within a short time. Fig. 4a shows one example of such base station switches. The user's phone switches to a base station nearly 6km away and then back to a nearby base station within 10 seconds.

Even when a user stays at the same place, his/her phone may switch among base stations nearby (Fig. 4b). Consequently, it is difficult to decide whether a user is actually moving or still.

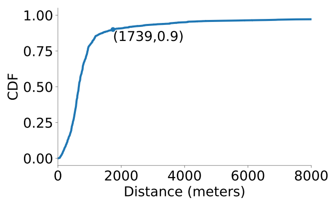
4.1.2 Bursty Sampling. Bursty sampling is another characteristic of MFRs. Mobile phone users usually access cellular network in a bursty and intermittent manner [16], *i.e.*, heavy data traffic within a short interval. For example, activities like watching online videos consume traffic intensively and continuously, causing a lot MFRs in a short time.

Fig. 4c illustrates the bursty sampling of MFRs. Points on each horizontal line represent the occurrences of a user's records in one day. Point $(t, user_i)$ means $user_i$ has a record at time t. Most users have one or two intervals of dense records separated by hours of blank except for user 3, who seems to be a heavy mobile phone user. One of User 6's "bursts" is zoomed in at the top of Fig. 4c. The records are sampled at a high frequency (from 0 times/min to 86 times/min, 7 times/min on average). The bursty sampling causes *redundancy* in the dense intervals, and lead to *sparsity* during blank intervals.

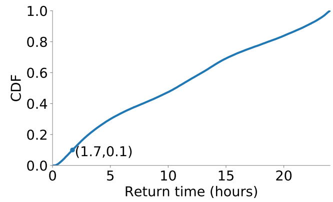
4.2 Base Station Switch Rectification

This subsection deals with the base station switch problem in MFRs. We treat switches to remote stations and switches to nearby stations differently.

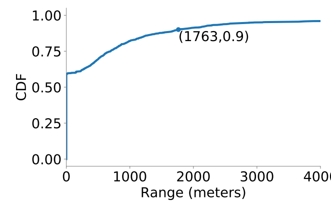
4.2.1 Switches to Remote Stations. Switches to remote stations can cause wrong location records in MFRs (Fig. 4a). We first sort each user's MFRs by time, and extract a sequence $\{p_i =< location, timestamp >=< p_i.loc, p_i.T >\}$, where $p_i.loc$ is the location of the base station that the phone connects to. Like [32]



(a)



(b)



(c)

Fig. 5. Cumulative Distribution Function of (a) distance between users and their connected base stations, (b) return time (The interval between users' leaving and coming back to the same place), (c) station switch range during stay.

dealing with station switch in CDR, we take a record p_i as a remote station switch if:

$$\begin{aligned} Dist(p_{i-1}.loc, p_i.loc) &> D_{noise} \\ p_i.T - p_{i-1}.T &< \Delta T_{noise} \end{aligned} \tag{1}$$

where $Dist$ is the Euclidian distance function. We do not use a speed threshold directly because nearby switches can also cause high speed as shown in Fig. 4b.

We set the threshold D_{noise} by analyzing the URI in MFRs (see Table 1). We observe that some location-based services embed users' GPS in the request URI, which can be seen as the actual locations of users. Fig. 5a shows the distance between users and their connected stations based on these records. For over 90% samples, the distance is below 1.8km, so we set D_{noise} to $2 \times 1.8 = 3.6$km. Considering the speed limit of a highway is 120km/h in China, we set the time threshold ΔT_{noise} to $3.6/120 \times 3600 = 108$s.

4.2.2 Switches to Nearby Stations. Nearby base station switches do not incur obviously wrong location records and can be considered as normal fluctuations of cellular localization. We propose techniques to extract stay points that are robust to variations of locations caused by nearby base station switches in Sec. 4.4.

4.3 Burst Separation

To handle the uneven sampling of MFRs, we divide the sequence of MFR logs of a user into multiple *bursty intervals* and *sparse intervals*, and process them differently when extracting stay points and durations (see Sec. 4.4).

A *bursty interval* is defined as $I_b =< p_1, p_2, ..., p_n >$, where

$$\begin{aligned} p_n.T - p_1.T &> \Delta T_{stay} \\ p_{i+1}.T - p_i.T &< \Delta T_{bursty} \ (i = 1, 2...n-1) \end{aligned} \tag{2}$$

Each interval between two neighbouring *bursty intervals* is a *sparse interval* (denoted by I_s).

To set ΔT_{bursty}, we consider that it should give a high probability that a user stays at the exact place during ΔT_{bursty} interval. Assume the gap between two consecutive records are ΔT, and the two records are generated at the same cellular tower. When ΔT is small enough ($\Delta T < \Delta T_{bursty}$), the user is very likely to stay near the cellular tower during this interval (ΔT). Otherwise, the user may have gone to another place and then come back. To determine the value of ΔT_{bursty}, we draw the distribution of the time between mobile phone users' two visits to the same location (*return* *time*). A *return* is identified by:

$$\begin{aligned} &< p_1, p_2, ..., p_i, ..., p_n > \\ &s.t.\ p_1.loc = p_n.loc, Dist(p_1, p_i) > D_{noise} \end{aligned} \tag{3}$$

Then *return time* is $p_n.T - p_1.T$. Based on Fig. 5b, *return time* is over 1.7 hours for 90% cases. For two successive records whose gap is shorter than 1.7 hours, we can infer that the user does not visit other places confidently. So ΔT_{bursty} is set to 1.7 hours. In Sec. 4.4 we will discuss the value of another threshold ΔT_{stay}.

In the inset of Fig. 4c, more than 82% of records are within 5 seconds after their predecessors. This means that the MFRs inside a bursty interval can be rather redundant. Most of records inside bursty interval can be discarded to accelerate data processing without losing information. Based on these observations, if a record is less than 10 seconds after the last record, it is considered redundant and removed. As a result, 69% of records are filtered out.

4.4 Stay Point Extraction

Based on MFR, we extract users *stay points* and then infer their homes and visits to commercial districts. When a mobile phone user stays in the neighbourhood for some time longer than a threshold, we call it a *stay point*. Stay points are a more robust way to represent users' location than raw location information directly from MFR, since phones can switch among nearby cellular towers even when they donot move. Stay points can also indicate semantic meanings about users' activities, such as visiting commercial districts and resting at home [39, 40].

We determine stay points as follows. First we split a user's MFRs into bursty intervals ($\{I_b\}$) and sparse intervals ($\{I_s\}$) as Sec. 4.3. Then stay points are extracted from each bursty interval $I_b =< p_1, p_2, ..., p_n >$. We define the neighbourhood of a record p_i as a circle centered at $p_i.loc$ with radius D_{nbh}.

Specifically, we first find the continuous records when a user stays in p_i's neighbourhood, *i.e.*,

$$\begin{aligned} &< p_s, ..., p_i, ..., p_e > \\ s.t.\ &Dist(p_j.loc, p_i.loc) <= D_{nbh}\ \forall s \le j \le e \\ &Dist(p_{s-1}.loc, p_i.loc) > D_{nbh} \\ &Dist(p_{e+1}.loc, p_i.loc) > D_{nbh} \end{aligned} \tag{4}$$

Then the time the user spends in p_i's neighbourhood is $p_i.st = p_e.T - p_s.T$. We select p_i with the maximum $p_i.st$ as p_{max}. If $p_{max}.st \ge \Delta T_{stay}$, then we extract a stay point $sp = (loc, arvT, levT)$:

$$\begin{aligned} &sp.loc = \sum_{k=s}^{e} p_k.loc/(e - s + 1) \\ &sp.arvT = p_s.T \quad sp.levT = p_e.T \end{aligned} \tag{5}$$

where $sp.loc$ is the center of the *stay point*, $sp.arvT$ and $sp.levT$ are the arrival and leaving time of *sp*, respectively.

After removing $I_{sp} =< p_s, ..., p_{max}, ..., p_e >$ from I_b, we update each remaining record's $p_i.st$ which are affected by the removal of I_{sp}. Then we repeat the above process to find other stay points until the maximum $p_i.st$ is shorter than ΔT_{stay}.

For sparse intervals, we only extract stay points at night and abandon the records at daytime. Note that the time between two records in sparse intervals can be larger than $\Delta T_{bursty} = 1.7$ hours, during which a *return* may occur (Fig. 5b). But if the sparse interval is at night, it is highly likely that consecutive records with the same location is a stay. This will help us extract home locations robustly.

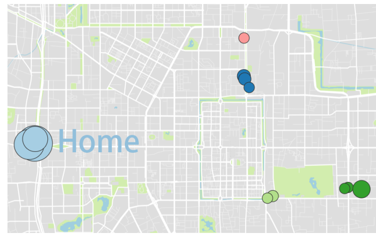


Fig. 6. Example of clustered stay points. Each circle is a stay point of the user. The area of a circle is proportional to its stay time at night. The color of the circle represents the cluster it belongs to.

The threshold ΔT_{stay} is the minimum time length of a stay. We set it to 20min because it suffices to qualify as a visit to commercial districts. Bursty intervals shorter than ΔT_{stay} will not contain stay points, so ΔT_{stay} is also used in Sec. 4.3 as the minimum length of bursty intervals.

We set the other threshold D_{nbh} by analyzing users' distribution of connected stations during stay. For GPS trajectories, D_{nbh} can be set manually to an appropriate value (200m [40]). But for MFRs, owing to the low spatial granularity and nearby station switches, the fluctuation range of connected stations is different from the range of users' wandering. From the records with GPS values (Sec. 4.2), we extract over 10 thousand stay points by the method of [40]. The distribution of station fluctuation range is shown in Fig. 5c. As is shown, about 60% stay points only have one station due to the low spatial granularity. Also, users' wandering can cause several kilometers of station fluctuation and 90% of them are below 1.763km. So D_{nbh} is set to 1.763km.

5 TRADE AREA DELINEATION

Based on the stay points extracted in Sec. 4.4, we infer the commercial districts' trade areas. First, we cluster Points of Interests (POI) to identify commercial districts automatically (Sec. 5.1) , then we infer the probabilities that residents visit each commercial district (Sec. 5.3), based on which we can quantify the trade areas.

5.1 Commercial District Identification

To find the commercial districts automatically, we adopt the method of [33]. We consider the POIs with annotations of commercial districts and shopping malls. First, the algorithm selects some seeds as the initial cluster centers, then assigns other POIs to their closest cluster unless the distance is greater than a threshold. By selecting a proper threshold (evaluated by the Silhouette Coefficient [26]), we obtain 52 commercial districts in the city of *D1* and 33 in *D2*.

5.2 Home Location Inference

In previous studies, researchers find that human mobility exhibits high regularity [13, 27], and human activities are usually centered around a few locations like home and work places [15]. To infer the probabilities that the residents in an area visit a commercial district, we first need to identify users' place of residence. We represent a user's stay points as $\{sp_1, sp_2, ..., sp_n\}$. After clustering with DBSCAN[11] (*epsilon* set to 500 meters and *minimum samples* set to 1), we find m clusters $\{c_1, c_2, ..., c_m\}$. For each cluster c_i:

$$
\begin{aligned}
c_i.st &= \sum_{sp \in c_i} (sp.levT - sp.arvT) \\
c_i.loc &= \sum_{sp \in c_i} sp.loc \times (sp.levT - sp.arvT)/c_i.st
\end{aligned}
\tag{6}
$$

where $c_i.st$ is the total stay time of the stay points in this cluster and $c_i.loc$ is the weighted centroid of stay points based on their stay time. Then we take the place where users stay most at night (20:00 to 8:00) as the users' home. Fig. 6 shows the clusters of a user's stay points, and the identified home is also marked. The stay points gather at a few key locations and the stay time at night of the home cluster is significantly larger than other clusters.

5.3 Visit Probability Estimation

First, We partition the city into 1km×1km grids (30km×30km totally). Then we infer the probabilities that people in each grid area visit each commercial district. If a user pays a visit to a commercial district between 18:00 and 23:00 on weekdays or 9:00 and 23:00 at weekends, which is the most common shopping time, it is counted as one visit to the commercial district. The visits are identified based on users' stay points, thus we can exclude the people that just pass by a commercial district. Besides, we take the place where users spend the most time at daytime (8:00 to 20:00) as their work locations, then we exclude the people working near a commercial district when counting the number of visits to this commercial district. P_{ij} is the probability that residents in area i visit commercial district j. It is calculated as $P_{ij} = C_{ij}/\sum_{k=1}^{N_i} C_{ik}$, where C_{ij} is the total number of visits from area i to district j and N_i is the total number of commercial districts. We admit that we are not able to differentiate people actually purchasing something from people purchasing nothing. The people visiting a commercial district without purchasing anything are potential customers for the commercial district. So understanding their behaviors is also beneficial to promote business profits.

6 TRADE AREA MODELING

This section investigates the impacting factors on trade areas of commercial districts based on the Huff model.

6.1 Basics on Huff Model

The Huff model [14] has been widely used for evaluating business geographic decisions including defining and analyzing trade areas. It models the visit probabilities from residential areas to commercial districts as below:

$$P_{ij} = \frac{U_{ij}}{\sum_{k=1}^{N_i} U_{ik}} \tag{7}$$

where P_{ij} is the probability that residents in area i visit commercial district j, N_i is the number of commercial districts, and U_{ij} is the utility of commercial district j to area i. Specifically,

$$U_{ij} = (\prod_{h=1}^{H} A_{hj}^{\gamma_h}) D_{ij}^{\lambda} \tag{8}$$

where A_{hj} is the h^{th} metric of the attractiveness of commercial district j and γ_h is the sensitivity parameter of P_{ij} to A_{hj}. D_{ij} is the distance (travel time) between area i and commercial district j with a negative sensitivity parameter λ to depict the distance decay effect.

We have calculated P_{ij} from MFRs in Sec. 5.3. The travel time D_{ij} can also be easily obtained via map services such as the *Baidu Map API* [1]. Below we describe how to determine the attractiveness A_{hj} and the sensitivity parameters.

6.2 Attractiveness Determination

The area is usually used to quantify attractiveness in previous works [31], while a consensus on the definition of attractiveness is currently absent. In this paper, we design various metrics from three categories of metrics to quantify the attractiveness of a commercial district in order to improve the accuracy of the Huff model.

6.2.1 Commercial Entity Metrics. The amounts and diversity of commercial entities in a district are important metrics that affect the attractiveness. For commercial district j, the numbers of shopping POIs (m_1), restaurant POIs (m_2) and entertaining POIs (m_3) are counted as the *commercial entity* metrics. To assess the diversity of entities, entropy measure (m_4) from information theory is applied to the frequency of commercial POI types.

6.2.2 Urban Facility Metrics. The attractiveness of a commercial district is not only related to commercial POIs, but also others like parking lots (m_5), scenic spots (m_6), bus stations (m_7), subway stations (m_8) and life services (m_9). They reflect the transportation accessibility and the services a district can provide. The numbers of these POIs are collected as the *urban facility* metrics.

6.2.3 Human Metrics. The population density and the incoming flow may have an impact on the trade area of a commercial district. Based on the locations of homes inferred from MFRs, we can estimate the population of an area. The population densities in 5km (m_{10}), 5~10km (m_{11}) and 10~15km (m_{12}) range around a commercial district are extracted. From MFR, we also get the incoming flow (m_{13}) for each commercial district, which excludes the residents in the commercial district.

6.3 Huff Model Fitting

Substitute Eq.(8) into Eq.(7), we get

$$P_{ij} = \frac{(\prod_{h=1}^{H} A_{hj}^{\gamma_h}) D_{ij}^{\lambda}}{\sum_{k=1}^{N_i} (\prod_{h=1}^{H} A_{hk}^{\gamma_h}) D_{ik}^{\lambda}} \tag{9}$$

Apply the following transformation, Eq.(9) can be transformed into a linear form:

$$\begin{aligned} log(\frac{P_{ij}}{\tilde{P_i}}) &= \sum_{h=1}^{H} \gamma_h \log \frac{A_{hj}}{\tilde{A_h}} + \lambda \log \frac{D_{ij}}{\tilde{D_i}} = W \cdot E \\ W &= (\gamma_1, ..., \gamma_H, \lambda) \\ E &= (\log \frac{A_{1j}}{\tilde{A_1}}, ..., \log \frac{A_{Hj}}{\tilde{A_H}}, \log \frac{D_{ij}}{\tilde{D_i}})^{\intercal} \end{aligned} \tag{10}$$

where $\tilde{P_i}$, $\tilde{A_h}$ and $\tilde{D_i}$ are respectively the geometric mean of P_{ij}, A_{hj} and D_{ij} over all commercial districts that residents in area i visited.

To automatically select the more relevant metrics of attractiveness, we apply $L^1 - norm$ to the solution of Eq.(10):

$$\hat{W} = \arg\min_{W} \{\beta \|W\|_1 + \frac{1}{2n} \|log(\frac{P_{ij}}{\tilde{P_i}}) - W \cdot E\|_2^2\} \tag{11}$$

where n is the number of samples and β is the weight of $L^1 - norm$. It has been shown that $L^1 - norm$ can bring sparsity to solutions that can be used to select effective metrics [33].

Once we obtain the value of W, we can analyze how much each metric contributes to the trade area of a commercial district (evaluated in Sec. 7.3.1), and predict the trade areas of other commercial districts (evaluated in Sec. 7.3.2).

7 EVALUATION

This section presents the evaluations of CellTradeMap. Due to the lack of ground truth on the actual locations of mobile phone users, it is difficult to evaluate the accuracy of the MFR processing

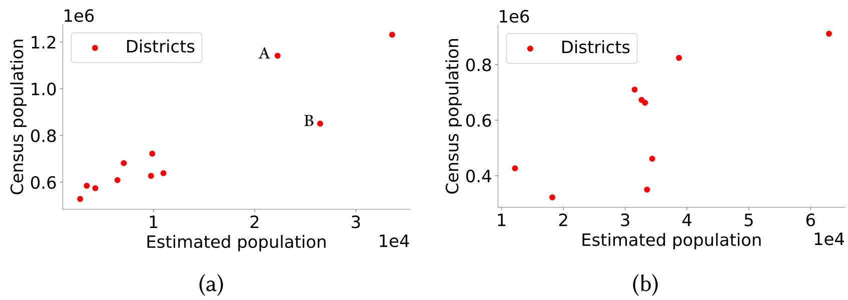


Fig. 7. Correlation between the number of residents inferred by CellTradeMap and that by census for each administrative district ((a): *D1*, (b): *D2*).

module. Hence we mainly assess the performance of CellTradeMap on trade area delineation and modeling.

7.1 Experimental Settings

We use the same MFR dataset as in Sec. 3.1. The POI data are from *Baidu Map API* [1].

We store the MFR dataset in a *Greenplum* [6] database, an open-source data platform for massively parallel processing. We use *d3.js*[7] and *mapbox* [8] to visualize trade areas. The remaining parts of the system are implemented in python and the experiments are run on a CentOS server with *Xeon E5* processor and 256 GB memory. The sampled data and code are available upon request to the corresponding author.

7.2 Performance of Trade Area Delineation

In this series of experiments, we evaluate the accuracy of CellTradeMap on home location inference and analyze the trade areas extracted from MFRs.

7.2.1 Accuracy of Home Location Inference. In this experiment, we compare the distribution of homes inferred by CellTradeMap with the census data published by the government for each administrative district. We evaluate the accuracy at the administrative district level rather than for each individual because we do not have access to the home information of each individual mobile phone user. To get robust results, we only use the users who have more than 4(20) days' records in *D1*(*D2*).

For *D1*, Fig. 7a plots the population of residents in each administrative district estimated by CellTradeMap (*i.e.*, whose homes are located in the district) and that obtained from governmental census data. We observe a strong linear correlation ($r = 0.90$) between the estimated population and the actual population in each administrative district. The only two outliers are district *A*, a suburban area, and district *B*, where the government resides. The deviation of these two points may be due to urbanization. The linear correlation implies almost unbiased sampling of residents among different administrative districts. The results of *D2* are shown in Fig. 7b. The inferred population is linearly related to the census data except for a few outliers (r=0.75).

7.2.2 Visualization of Trade Areas. In this experiment, we calculate the visit probabilities of residents to each commercial district, and plot the *(i)* contour maps of visit probabilities and *(ii)* heatmaps of visitors to get insights on the trade areas.

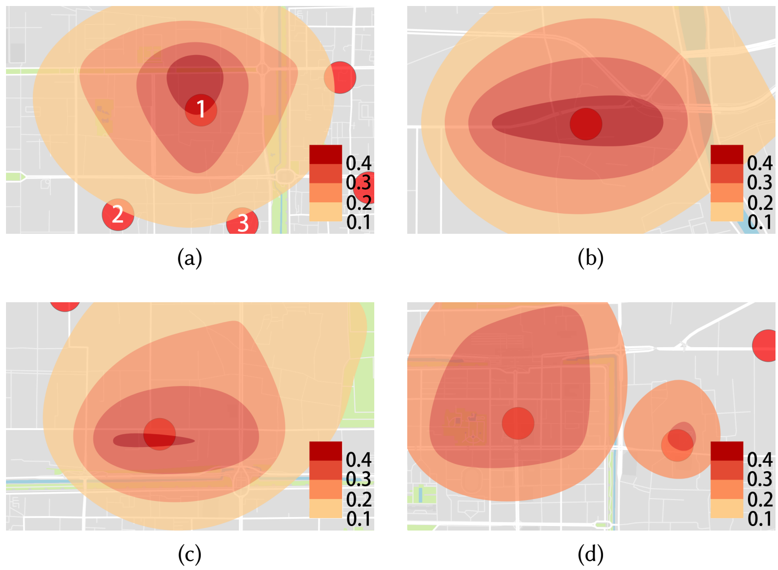
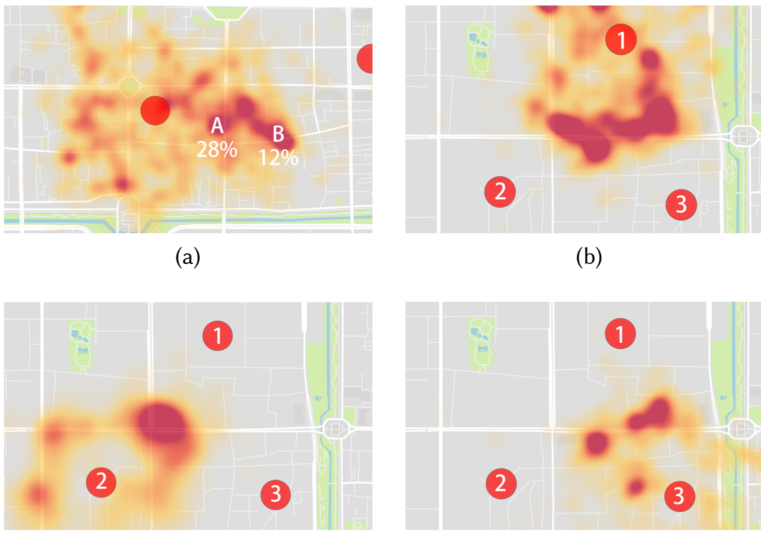


Fig. 8. Contour maps of visit probabilities in *D1*. Circle nodes represent the center of commercial districts. The color of an area reflects the probability that residents in this area visit a specific commercial district. The probability is calculated in Sec. 5.3. (a) The trade area of 1 is squeezed in the south due to the competition from district 2 and 3, but stretched in the north due to the east-west road. (b) There is no competition for this commercial district. The trade area extends nearly uniformly except for the stretch along the east-west road. (c) The extension of the trade area to the south is blocked by a river. (d) The trade area of district 1 is much larger than that of district 2 owing to the different attractiveness of the two districts.

Fig. 8 and Fig. 10(a)(b) show representative contour maps of visit probabilities in *D1* and *D2*. First, we calculate the visit probabilities from grids to commercial districts as in Sec. 5.3, then we take these probabilities as the samples at the center of each grid, finally we get the contour lines based on these samples [3]. In these figures, The color of an area reflects the probability that residents in this area visit a specific commercial district.

We obtain the following insights from the different patterns of trade areas.

(1) The *competition* from nearby commercial districts can compress the trade area. For example, in Fig. 8a, the trade area of commercial district 1 is squeezed by the competition with districts 2, 3, which means that the market share of commercial district 1 in the central area is decreased. In Fig. 10a, the trade area is squeezed horizontally by the competition from nearby competitors.
(2) The *road network* is another reason for the anisotropy of the trade area. In Fig. 8b, due to the east-west road passing by, the trade area elongates along the road. Except for this, the trade area extends almost evenly because there are no other commercial districts nearby.
(3) The *natural barriers* like rivers can cut off the spread of the trade area. As shown in Fig. 8c, a river lying in the south blocks the residents on the south bank to visit the commercial district on the north bank, whose trade area spread much further to the north. In Fig. 10b, the extension of trade area to the southeast is also blocked by a river.
(4) The *attractiveness* may lead to different sizes of trade areas. As shown in Fig. 8d, the two closely located commercial districts have different sizes of trade areas.



(a) (b) (c) (d)

Fig. 9. Heatmaps showing the distribution of visitors' homes in *D1*. Circle nodes represent the center of commercial districts. The intensity of red represents the absolute number of visits from a location. (a) *A* and *B* are both major sources of visitors, but the visit probabilities for residents in *A* and *B* to visit this commercial district are different. (b), (c) and (d) illustrate the same three commercial districts with Fig. 8a. (b), (c) and (d) show the distribution of visitors of commercial districts 1, 2, 3 respectively.

Fig. 9, Fig. 10c and Fig. 10d show heatmaps of commercial districts' visitors in *D1* and *D2*. The intensity of color represents the number of visitors from this area.

(1) In Fig. 9a, location *A* and *B* are two major sources of visitors for the commercial district, but the market shares at these two locations differ, 28% at *A*, while 12% at *B*.
(2) Fig. 9b, Fig. 9c and Fig. 9d illustrate the distribution of visitors for the three commercial districts in Fig. 8a. We find that the middle area among the three commercial districts is a major source of visitors for all the three districts, although the visit probability to each district is relatively low owing to the competition. Such areas with low market share and large volume of visitors should be the focus of business managers.
(3) The visitors in Fig. 10c mostly come from the left of the commercial district while the visitors in Fig. 10d come from a much broader area.

7.3 Performance of Trade Area Modeling

In this series of experiments, we identify the key metrics of attractiveness and assess the accuracy of the Huff model fitted by CellTradeMap to predict the trade areas of other commercial districts using 5-fold cross validation. Specifically, the commercial districts are divided randomly and evenly into 5 groups. In each round of cross validation, one group is used for testing and the other four are used for training.

7.3.1 Sensitivity Analysis of Attractiveness Metrics. In this experiment, the sensitivity parameters $\gamma_1, \gamma_2, ..., \gamma_H$ are solved from Eq.(10) and each parameter corresponds to a metric of attractiveness.

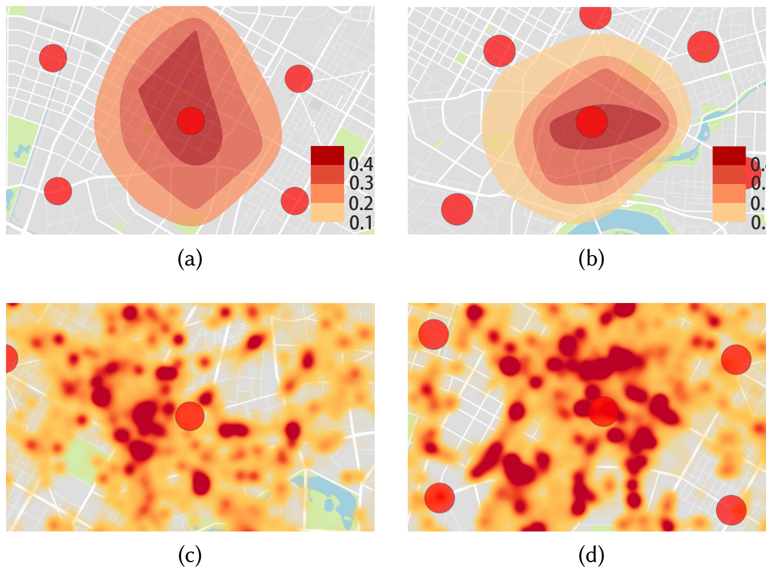


Fig. 10. *D2*: (a)(b) Contour maps of visit probabilities. (c)(d) Heatmaps showing the distribution of visitors' homes who have visited the commercial district at the center of the figure. (a) The trade area is squeezed horizontally by the competition from nearby competitors. (b) The extension of trade area to the southeast is blocked by a river. (c) Most visitors come from the left of the commercial district. (d) The commercial district at the center attracts visitors from a broad area.

The sensitivities are averaged over 5-fold cross validation and the metrics with top sensitivity are shown in Fig. 11a and Fig. 11b:

(1) In the city of *D1*, abundant restaurant options and parking lots, and large crowd flows are critical to the attractiveness of a commercial district. Besides, easy access to public transportation and having scenic spots are also helpful.
(2) In the city of *D2*, the number of restaurants and crowd flows are key metrics of commercial attractiveness, while having more shops, life services and entertainment entities are also helpful.

The results show that the attractiveness metrics have both similarities and differences in different cities. Large crowd flows and abundant restaurant choices are key metrics in both cities, while the other metrics that help improve attractiveness are different. The city of *D1* is a famous tourism city, thus having scenic spots is an important metric. The other dissimilarities may be related to the different life styles in two cities.

It should also be noted that the ubiquitous coverage of MFR is important for sensitivity analysis. Fig. 12a and Fig. 12b show the sensitivity analysis based on 5 randomly sampled commercial districts. Compared to Fig. 11a and Fig. 11b, the variances (error bars) are much larger for the sampled 5 districts, which means that sensitivity analysis with a small number of commercial districts tends to be unreliable.

7.3.2 Predictive Accuracy of Trade Area Model. In this experiment, we utilize the Huff model fitted using commercial districts in the training set to predict the visit probabilities P_{ij} of commercial

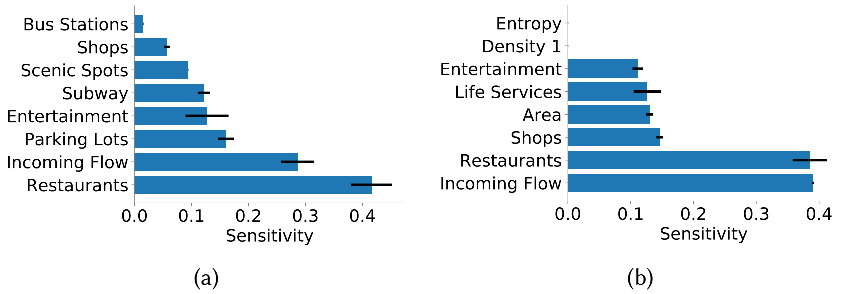


Fig. 11. The top 8 attractiveness metrics with high sensitivities. The error bar is the variance over 5-fold cross-validation. Density 1, 2 and 3 are the population densities in 5km, 5~10km and 10~15km range around a commercial district respectively. ((a): *D1*, (b): *D2*)

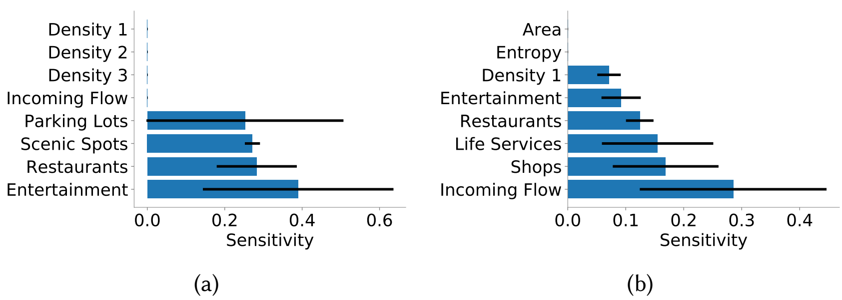


Fig. 12. The top 8 attractiveness metrics obtained from 5 commercial districts. Density 1, 2 and 3 are the population densities in 5km, 5~10km and 10~15km range around a commercial district respectively. ((a): *D1*, (b): *D2*)

districts in the testing set. The accuracy is measured by the root mean square error (RMSE) of P_{ij}:

$$RMSE = \sqrt{\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(P_{ij} - \hat{P_{ij}})^2} \tag{12}$$

where I, J are the numbers of residential areas and commercial districts. $\hat{P_{ij}}$ is the estimated P_{ij}.

We compare CellTradeMap with two baselines.

- *Linear Regression.* Least squares method is used to calibrate the Huff model with all 13 metrics.
- *Random.* Linear Regression with 5 randomly selected metrics to calibrate the Huff model.

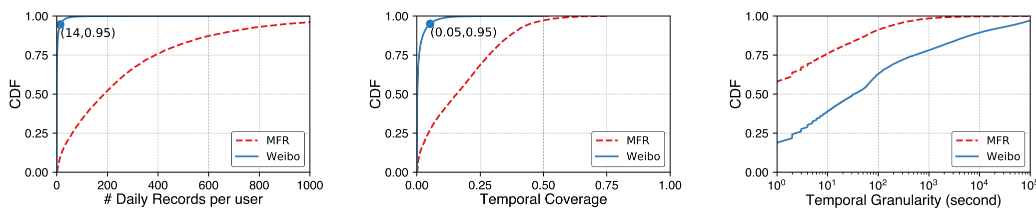
Table 3 and Table 4 summarize the results from 5-fold cross validation in *D1* and *D2*. The model fitted by CellTradeMap yields the best RMSE in both cities. *Linear Regression* performs worst, since too many irrelevant metrics will harm the model's accuracy. Compared with *Random*, the decrease of RMSE implies that with the help of $L^1 - norm$, CellTradeMap can improve the accuracy by selecting the most important attractiveness metrics like *Incoming Flow* based on MFR.

Table 3. D1, Average RMSE on prediction accuracy.

Method	RMSE
Linear Regression	0.188
Random_5	0.145
$L^1 - norm$	0.128

Table 4. D2, Average RMSE on prediction accuracy.

Method	RMSE
Linear Regression	0.160
Random_5	0.155
$L^1 - norm$	0.148



(a) (b) (c)

Fig. 13. Compare MFR and Weibo data, which is a superset of check-in data. (a) CDF of the averaged number of records per user per day. (b) CDF of temporal coverage. (c) CDF of inter-record intervals.

7.4 Comparison of MFR and Check-in

We do not have check-in data, but we have the application information in MFRs. Weibo is the application with most check-ins in China. We retrieve all the MFRs of Weibo as a superset of the check-in data. We compare MFR and the superset of check-in from four aspects: daily records per user, temporal coverage, temporal granularity and home inference. We do not compare their performance in trade area analysis directly due to the lack of groundtruth.

- **Daily records per user:** Fig. 13a shows the CDF of the averaged number of records per user per day. Most users have less than 10 Weibo records per day and a part of these records correspond to users' check-in. The network traffic of Weibo is only a part of the overall traffic logged by MFR, not to mention the check-in data. So it is reasonable that there are much more MFRs than check-ins.
- **Temporal coverage:** In Sec. 4.1.2, we discussed the redundancy in MFR. So more data do not necessarily provide more information about users' locations. We segment one day into 48 intervals uniformly. Then we count the number of intervals that MFR and Weibo data cover. The *temporal coverage* is defined as:

$$temporal\_coverage = \frac{\text{\# of intervals covered}}{48}$$

  The results are shown in Fig. 13b. For most users, Weibo data cover a very small fraction of their daily activities (less than five percent). However, MFRs cover a much larger fraction which makes MFRs more suitable to analyze users' home locations and shopping activities.
- **Temporal granularity:** Fig. 13c shows the CDF of inter-record intervals. Most intervals of MFR are shorter than 10^3 s (about 17 min), while many consecutive Weibo records are hours apart. The check-in data would be even sparser. This makes check-in data very easily to miss users' activities like visiting commercial districts.
- **Home inference:** In Sec. 7.2.1, we evaluate the accuracy of home location inference by the correlation analysis between the inferred number of residents and the census data. The

Table 5. Correlation Coefficient of Home Location Inference

Data	Correlation coefficient
MFR	0.75
Weibo	0.58

correlation coefficient measures how strong two variables are linearly correlated, thus it can indicate the accuracy of home location inference since we do not have groundtruth for individual home location. For *D2*, we infer users' homes with MFR or Weibo data separately and calculate the pearson correlation coefficient. The results are shown in Table 5. The lower coefficient of Weibo data suggests that Weibo data is biased among different administrative districts.

Compared to MFR, check-in data are sparse and biased, which makes it unable to support precise and urban scale trade area analysis.

8 CONCLUSION AND FUTURE WORK

In this paper, we propose CellTradeMap, a novel cellular network-based trade area analysis framework for commercial districts. We devise processing techniques to extract robust location information from flow-level cellular data, and design analytical methods to adapt the conventional trade area analysis workflow to integrate cellular data. We evaluate the performance of CellTradeMap on trade area delineation and modeling using an urban-scale cellular network dataset covering 3.5 million mobile phone users. Experimental results show that CellTradeMap is able to extract explainable trade areas, identify important attractiveness metrics, and predict trade areas of an unseen commercial district with high accuracy. We envision our work as a pilot study to unlock the full business potentials of big cellular data analysis.

Looking forward, we will investigate the practical deployment of CellTradeMap. For example, how many records do we need to achieve reliable results, and how to determine the tradeoff between the data coverage and the system overhead. Another important direction is the generalization of the results of CellTradeMap, whether the results based on the data of one city can be generalized to other cities.

9 ACKNOWLEDGMENTS

This work is supported in part by the National Key Research Plan under grant No. 2016YFC0700100, the NSFC under grant 61832010, 61632008, 61672319, 61872081, 61632013, and Microsoft Research Asia.

REFERENCES

[1] Baidu. [n.d.]. *Baidu Location Based Services.* Retrieved March 11, 2019 from http://lbsyun.baidu.com/
[2] Zachary Davies Boren. 2014. There are officially more mobile devices than people in the world. *The Independent* 7 (2014).
[3] Mike Bostock. [n.d.]. *d3-contour.* Retrieved March 11, 2019 from https://github.com/d3/d3-contour
[4] Francesco Calabrese, Laura Ferrari, and Vincent D Blondel. 2015. Urban sensing using mobile phone network data: a survey of research. *Comput. Surveys* 47, 2 (2015), 25.
[5] Cisco. 2017. Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper.
[6] Greenplum Database. [n.d.]. *Massively Parallel Postgres for Analytics.* Retrieved March 11, 2019 from https://greenplum.org/
[7] D3 development group. [n.d.]. *Data-Driven Documents.* Retrieved March 11, 2019 from https://d3js.org/

[8] Mapbox development group. [n.d.]. *Mapbox: Explore. Move. Connect. Build with Live Location Data.* Retrieved March 11, 2019 from https://www.mapbox.com

[9] Subhankar Dhar and Upkar Varshney. 2011. Challenges and business models for mobile location-based services and advertising. *Commun. ACM* 54, 5 (2011), 121–128.

[10] Ela Dramowicz. 2005. Retail trade area analysis using the Huff model. *Directions Magazine* 2 (2005).

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *KDD*. ACM, 226–231.

[12] Laura Ferrari, Marco Mamei, and Massimo Colonna. 2012. People get together on special events: Discovering happenings in the city via cell network analysis. In *PerCom Workshops*. IEEE, 223–228.

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

[14] David L Huff. 1963. A probabilistic analysis of shopping center trade areas. *Land Economics* 39, 1 (1963), 81–90.

[15] 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*. Springer, 133–151.

[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 *CellNet*. ACM, 7–12.

[17] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013. Geo-spotting: mining online location-based services for optimal retail store placement. In *KDD*. ACM, 793–801.

[18] Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. 2011. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *CHI*. ACM, 2409–2418.

[19] 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.

[20] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. 2016. Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials* 18, 1 (2016), 124–161.

[21] Robert A Peterson. 1974. Trade area analysis using trend surface mapping. *Journal of Marketing Research* 11, 3 (1974), 338–342.

[22] Ling Qi, Yuanyuan Qiao, Fehmi Ben Abdesslem, Zhanyu Ma, and Jie Yang. 2016. Oscillation resolution for massive cell phone traffic data. In *MOBIDATA*. ACM, 25–30.

[23] Yan Qu and Jun Zhang. 2013. Trade area analysis using user generated mobile location data. In *WWW*. ACM, 1053–1064.

[24] William John Reilly. 1929. Method for the Study of Retail Relationships. *University of Texas Bulletin* 2944 (1929).

[25] William John Reilly. 1931. *The law of retail gravitation.* WJ Reilly.

[26] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.

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

[28] Etienne Thuillier, Laurent Moalic, Sid Lamrous, and Alexandre Caminada. 2018. Clustering weekly patterns of human mobility through mobile phone data. *IEEE Transactions on Mobile Computing* 17, 4 (2018), 817–830.

[29] 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 *IMC*. ACM, 225–238.

[30] Xu Wang, Zimu Zhou, Fu Xiao, Kai Xing, Zheng Yang, Yunhao Liu, and Chunyi Peng. 2018. Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Transactions on Mobile Computing* (2018).

[31] Yandong Wang, Wei Jiang, Senbao Liu, Xinyue Ye, and Teng Wang. 2016. Evaluating trade areas using social media data with a calibrated huff model. *ISPRS International Journal of Geo-Information* 5, 7 (2016), 112.

[32] Peiyu Yang, Tongyu Zhu, Xuejin Wan, and Xuejiao Wang. 2014. Identifying significant places using multi-day call detail records. In *ICTAI*. IEEE, 360–366.

[33] Su Yang, Minjie Wang, Wenshan Wang, Yi Sun, Jun Gao, Weishan Zhang, and Jiulong Zhang. 2017. Predicting Commercial Activeness over Urban Big Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 119.

[34] 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 *MobiCom*. ACM, 201–212.

[35] 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.

[36] Xinglin Zhang, Zheng Yang, Zimu Zhou, Haibin Cai, Lei Chen, and Xiangyang Li. 2014. Free market of crowdsourcing: Incentive mechanism design for mobile sensing. *IEEE transactions on parallel and distributed systems* 25, 12 (2014), 3190–3200.

[37] Ying Zhang. 2014. User mobility from the view of cellular data networks. In *INFOCOM*. IEEE, 1348–1356.

[38] 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 *INFOCOM*. IEEE.

[39] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. 2010. Collaborative location and activity recommendations with GPS history data. In *WWW*. ACM, 1029–1038.

[40] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *WWW*. ACM, 791–800.