CellTradeMap: Delineating Trade Areas for Urban Commercial Districts with Cellular Networks

Yi Zhao*, Zimu Zhou†, Xu Wang*, Tongtong Liu*, Yunhao Liu*‡, Zheng Yang*
*School of Software and BNRist, Tsinghua University
†Computer Engineering and Networks Laboratory, ETH Zurich
‡Department of Computer Science and Engineering, Michigan State University
zhaoyi.yuan31@gmail.com, zzhou@tik.ee.ethz.ch, {darenwang11, liutongtong7, yunhaoliu, hmilyyz}@gmail.com

***Abstract*—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. 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 a large-scale cellular network dataset covering 3.5 million mobile phone users in a metropolis in China. 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.**

I. 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 worldwide [1]. These mobile devices frequently access the Internet via 3G/4G/5G networks for various applications including 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 [2].

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 [3], [4]. Over the past decade, researchers have exploited cellular data to mine customer mobility behaviour for various business strategies and applications such as mobile advertising [5], optimal store location planning [6] and commercial activeness prediction [7].

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 [8]. 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.

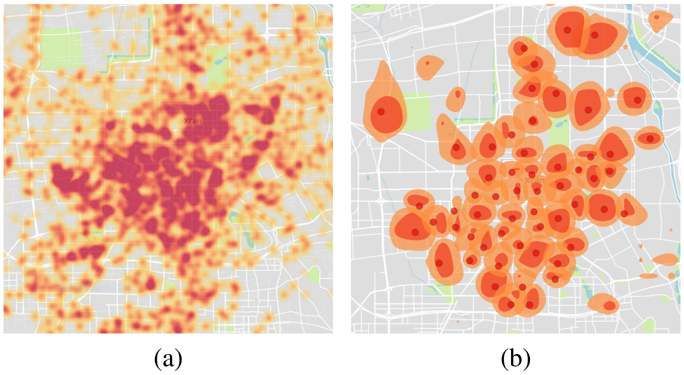

(a) (b)

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.

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. Response rates are also relatively low [9].

Pioneer studies [6], [9], [10] propose to utilize location-based social media as an alternative data source. Check-ins on social media are easier to collect at large-scale than surveys, and prove effective in profiling trade areas of popular retail stores [10]. However, it is difficult to infer place of residence and obtain comprehensive visitation information of competitor businesses from the limited and biased check-in data [11]. Furthermore, it is difficult to aggregate the trade area of stores to obtain the trade area of commercial districts without bias.

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 [12]. 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. [10]).

We propose CellTradeMap, a MFR-based framework to delineate and model trade areas for commercial districts. We base our design upon a large-scale MFR dataset covering 3.5 million anonymized mobile phone users in a metropolis of China within one week. 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 [13] to incorporate this cellular data, and improve the accuracy of the widely adopted trade area model [8] 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.

In the rest of this paper, we review related work in Sec. II, introduce our dataset and CellTradeMap framework in Sec. III, present the details of the three modules of CellTradeMap in Sec. IV, Sec. V and Sec. VI, and evaluate its performance in Sec. VII. Finally we conclude this paper in Sec. VIII.

II. 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.

A. Urban Sensing with Cellular Networks

Due to the deep penetration and the ubiquitous coverage, cellular networks are ideal for large-scale and comprehensive urban sensing [12], [14], [15]. Researchers have exploited different types of cellular data for various urban sensing applications.

Aggregated cellular network traffic has been widely applied to urban cellular traffic monitoring and management. Ferrari *et al.* [16] divide the city into grids and aggregate cellular usage data in each grid to detect special events. Wang *et al.* [17] aggregate cellular network usage data at the cell tower level to predict future cellular traffic in the city.

Call detail record (CDR) data records a time stamp and the connected cell ID in case of a phone call or a message. CDR data contains information about the long-term mobility and have been used in studying the patterns and fundamental laws of human mobility [3], [18], [19]. CDR data can also be integrated with public transit data to derive more accurate human mobility patterns [20].

MFR data are system logs that are sampled whenever a user accesses the Internet. Previous studies have harnessed MFR for fine-grained cellular traffic characterization [21] and human mobility modeling [4], [22]. 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.

B. Trade Area Analysis

Trade area analysis is an urban sensing application that answers questions such as "how faraway did the customers travel" and "what are the impacting factors to attract customers" to a store or a commercial district. These questions are important for city planning and understanding consumer behaviour [23]. Essential in trade area analysis is how to estimate the number of visitation to stores or commercial districts. Traditional methods [24], [25] use surveys to estimate the number of visits.

User check-ins on social networks emerge as a low-cost alternative to estimate the number of visitation [6], [9], [10]. Wang *et al.* [10] 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.* [9] 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 [10], 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.

III. Overview

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

A. 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 host and the Uniform Resource Identifier (URI) of the request, as well as other flow information like upload/download bytes and round-trip time (o2r/r2o and rtt in

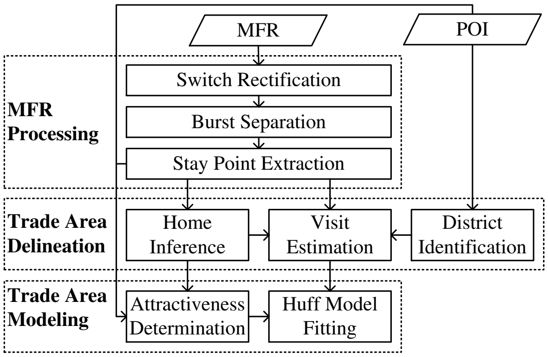


Fig. 2. Overview of CellTradeMap.

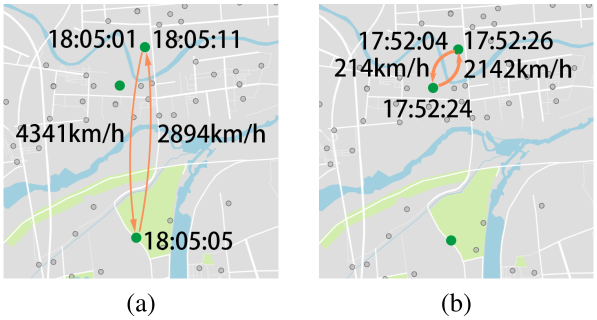


Fig. 3. Switches to (a) a remote base station and (b) a nearby base station. The green (bigger) dots are the user's connected stations at the corresponding time, and the grey (smaller) dots are all stations nearby.

Table I). Our MFR dataset is recorded by a major carrier in a Chinese metropolis, covering an area about 10 thousand km^2 over one week and our analysis focuses on the $30\text{km} \times 30\text{km}$ urban area. The dataset contains 17 billion anonymized MFRs from 3.5 million mobile phone users.

The MFR dataset covers a wide spatial range and has a high time resolution. In comparison, the check-in dataset used in [10] 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 [10] is only 0.1. Similarly, the average interval between consecutive records of one user in the CDR dataset used by [19] is 8.2 hours.

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.

B. 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. 2).

- **MFR Processing.** This module extracts *stay points and durations* of mobile phone users from raw MFRs. Stay points and durations are the basis to identify visits to commercial districts and locations of residence, which are traditionally obtained by expensive and time-consuming surveys. Recent proposals exploit check-ins from social media to count visitations [6], but such data is prone to sparsity and bias [10]. Techniques to extract stay points from GPS traces [26], [27] cannot be applied to MFRs because of MFRs' unique characteristics (Sec. IV-A). 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. IV.
- **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. V). We also explain the different patterns of trade areas (Sec. VII-B2).
- **Trade Area Modeling.** This module associates contexts such as the attractiveness of a commercial district to its visit probability. The Huff gravity model [8] 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. VI).

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

IV. Mobile Flow Record Processing

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

A. Challenges

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 [28]. Sometimes a phone may suddenly connect to a remote base station, exchange several packets and switch back within a short time. Fig. 3a 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. 3b). Consequently, it is difficult to decide whether a user is actually moving or still.

2) Bursty Sampling: Another characteristic of MFRs is their bursty sampling. This is because mobile phone users often access data services in a bursty and intermittent way [29], *i.e.*, intensive data usage within a short time. For example, mobile Internet access activities such as watching online videos usually consume cellular data intensively and continuously, resulting in a large volume of MFRs. However, unless users are addicted to their phones, these heavy-traffic activities tend to be separated by intervals with few Internet accesses.

TABLE I
EXAMPLE OF MOBILE FLOW RECORD.

User	Time	Station	Host	URI	o2r Bytes	r2o Bytes	RTT	Open Reason	Close Reason	...
$user_1$	t_1	s_1	www.example.com	/index/...	614	418	53	tcpSyn	finClose	...
...	...	...	...	...	...	...	...	...	...	...

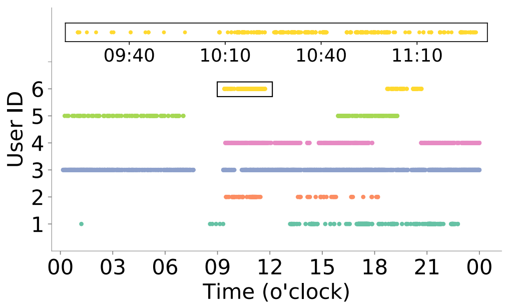


Fig. 4. Bursty sampling of MFRs. Points on each horizontal line represent the occurrences of one user's MFRs. One of User 6's bursts is zoomed in at the top.

Fig. 4 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. 4. 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.

B. 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.

1) Switches to Remote Stations: Switches to remote stations can cause wrong location records in MFRs (Fig. 3a). They are harmful to stay point extraction because they may split a stay into fragments, which do not qualify as a "stay" individually. 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 [30] 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. 3b.

We set the threshold D_{noise} by analyzing the URI in MFRs (see Table I). 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 $2 \times 1.8 = 3.6\text{km}$. Considering the speed limit of a highway is 120km/h in China, we set the time threshold ΔT_{noise} to $3.6/120 \times 3600 = 108\text{s}$.

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. IV-D.

C. 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. IV-D).

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).

ΔT_{bursty} is set such that a user stays at the same location during ΔT_{bursty} with a high probability. We study the distribution of users' *return time*, *i.e.*, the time between two visits to the same place. A *return* is identified by:

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

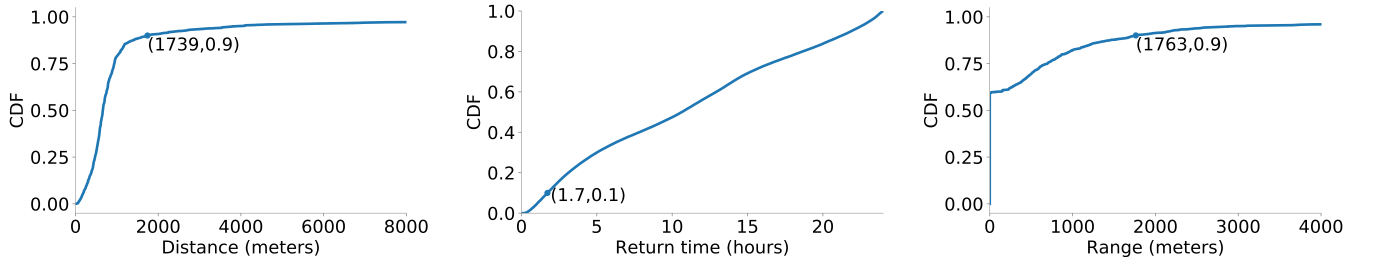
Then *return time* is $p_n.T - p_1.T$. As shown in Fig. 5b, 90% of *return time* is over 1.7 hours. If there are two consecutive records with the same location and the interval is below 1.7 hours, we can infer that the user stays at the same location with high confidence. So we set ΔT_{bursty} to 1.7 hours.

The other threshold ΔT_{stay} is the minimum length of a bursty interval. We set it to 20 min and defer the details to stay point extraction (Sec. IV-D).

Note that the MFRs within a bursty interval can be redundant. For example, over 82% of consecutive records in the inset of Fig. 4 are within 5 seconds. Due to the speed limit of humans and the continuity of movement, these records contain the same location information of the user and can be ignored to speed up data processing. A record is considered redundant if it is within 10 seconds with its predecessor and has the same location. As a result, 69% of records are filtered out.

D. Stay Point Extraction

We extract *stay points* from users' MFRs to identify users' homes and their visits to commercial districts. A stay point



(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.

approximates the location of a mobile phone user when he/she stays in a *geographic neighbourhood* over a time threshold. A stay point is a more robust estimate of a user's location than the location information contained in each MFR, because a mobile phone can switch among nearby base stations even when the user stands still. Another advantage of using stay points is that they indicate semantic meanings such as visiting commercial districts, resting at home and working at company [26], [27].

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. IV-C. 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.*,

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

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 $p_{max}.st \geq \Delta T_{stay}$, 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.

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. IV-C 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 [27]). 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. IV-B), we extract over 10 thousand stay points by the method of [27]. 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.

V. Trade Area Delineation

In this section, we utilize the stay points and durations extracted in Sec. IV-D to infer the trade areas of commercial districts in the city. The trade area of a commercial district can be quantified by the visit probabilities of residents from different areas. We explain how to identify commercial districts from POIs (Sec. V-A) and infer home locations from MFRs (Sec. V-B). Then we can easily calculate the probabilities that residents visit commercial districts (Sec. V-C).

A. Commercial District Identification

We use the clustering algorithm proposed by [7] to aggregate the POIs with annotations of shopping malls or commercial streets into commercial districts. First, different seeds are selected as centers to initialize the clusters. Then other POIs are gradually assigned to their closest cluster unless the distance is greater than a threshold. By selecting a proper threshold (evaluated by the Silhouette Coefficient [31]), we obtain 52 commercial districts. Each commercial district covers the convex hull of the POIs in the cluster.

B. Home Location Inference

Previous studies on human mobility indicate that people's movement exhibits high regularity [3], [19], and people's activities are usually around a few key locations such as home and work [32]. Instead of clustering a user's raw location records directly [10], we cluster a user's stay points to identify the important places (like home) in his/her lives, because stay points are more robust and contain more semantic meanings like resting at home. The clustering algorithm we adopted is DBSCAN [33]. Assume a user's stay points are $\{sp_1, sp_2, ..., sp_n\}$ and after clustering, 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 sum of stay time of all stay points in the cluster and $c_i.loc$ is the centroid of all stay points in the same cluster weighted by their stay durations. Then the cluster with longest stay time at night (20:00 to 8:00) is considered the home location of the user.

C. Visit Probability Estimation

We divide the city into $1\text{km} \times 1\text{km}$ areas ($30\text{km} \times 30\text{km}$ totally) and aggregate the homes of mobile phone users to each area. A stay point in a commercial district is counted as a visit if it is between 18:00 and 23:00 on weekdays or 9:00 and 23:00 at weekends, *i.e.*, the most common shopping time (The time window can be adjusted due to different analysis requirements). The probability P_{ij} for residents in area i to visit commercial district j is calculated as $P_{ij} = C_{ij} / \sum_{k=1}^{N_i} C_{ik}$, where C_{ij} is the number of visits from area i to district j and N_i is the number of commercial districts. We can then plot the contour maps of visit probabilities or heatmaps of visitors' homes to visualize the trade areas of commercial districts (Sec. VII-B2).

VI. Trade Area Modeling

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

A. Basics on Huff Model

The Huff model [8] 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. V-C. The travel time D_{ij} can also be easily obtained via map services such as the *Baidu Map API*. Below we describe how to determine the attractiveness A_{hj} and the sensitivity parameters.

B. Attractiveness Determination

There is no consensus on the definition of attractiveness in the Huff model. By default, the attractiveness is set to be the square footage [9]. To improve the accuracy of the Huff model, we propose three categories of metrics to quantify the attractiveness of a commercial district.

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. The higher the entropy is, the more heterogeneous the commercial entities are.

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.

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.

All the three categories of metrics are aggregated into a vector to represent the attractiveness of a commercial district: $(A_{1j}, A_{2j}, ..., A_{Hj}) = (m_1, m_2, ..., m_{13})$ As we will show in Sec. VII-C1, human metrics such as incoming flows are vital to quantify the attractiveness, which are difficult to obtain without using fine-grained cellular network data such as MFRs.

C. Huff Model Fitting

Substitute Eq.(8) into Eq.(7) and apply the following transformation, Eq.(7) can be transformed into a linear form:

$$\begin{aligned} log(\frac{P_{ij}}{\tilde{P}_i}) &= \sum_{h=1}^{H} \gamma_h \log \frac{A_{hj}}{\tilde{A}_h} + \lambda \log \frac{D_{ij}}{\tilde{D}_i} = W \cdot E \\ W &= (\gamma_1, ..., \gamma_H, \lambda) \\ E &= (\log \frac{A_{1j}}{\tilde{A}_1}, ..., \log \frac{A_{Hj}}{\tilde{A}_H}, \log \frac{D_{ij}}{\tilde{D}_i})^{\intercal} \end{aligned} \tag{9}$$

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.(9):

$$\hat{W} = \arg\min_{W}\{\beta\|W\|_1 + \frac{1}{2n}\|log(\frac{P_{ij}}{\tilde{P}_i}) - W \cdot E\|_2^2\} \quad (10)$$

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 [7].

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. VII-C1), and predict the trade areas of other commercial districts (evaluated in Sec. VII-C2).

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

A. Experimental Settings

We use the same MFR dataset as in Sec. III-A. Our POI dataset is collected through *Baidu Map API*, which contains 560 thousand POIs.

The MFR dataset is stored in a *Greenplum* database, which is an open-source and massively parallel processing data platform. The trade area visualization part is implemented with *d3.js* and *mapbox*. The other parts of CellTradeMap are implemented in Python3.6, running on a CentOS 6.8 server with *Xeon E5* processor and 256 GB memory. The sampled data and code are available upon request to the corresponding author.

B. 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.

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 days' records.

Fig. 6a 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. Consequently, the trade areas delineated by CellTradeMap tend to be comprehensive and unbiased.

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. 7 shows four representative contour maps of visit probabilities. In these figures, The color of an area reflects the probability that residents in this area visit a specific commercial district, which is calculated in Sec. V-C. We obtain the following insights from the different patterns of trade areas.

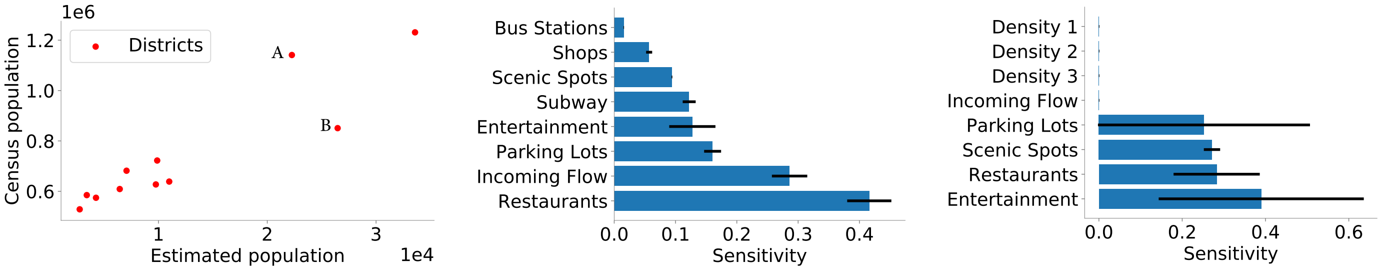
1) The *competition* from nearby commercial districts can compress the trade area. For example, in Fig. 7a, the trade area of commercial district 1 is squeezed by the competition with districts 2, 3, which means that the market share of commercial district 1 in the central area is decreased.
2) The *road network* is another reason for the anisotropy of the trade area. In Fig. 7b, 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. 7c, 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.
4) The *attractiveness* may lead to different sizes of trade areas. As shown in Fig. 7d, the two closely located commercial districts have different sizes of trade areas.

Fig. 8 further shows four representative heatmaps of visitors. The intensity of color represents the number of visitors from this area. In Fig. 8a, 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. Fig. 8b, Fig. 8c and Fig. 8d illustrate the distribution of visitors for the three commercial districts in Fig. 7a. 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.

C. Performance of Trade Area Modeling

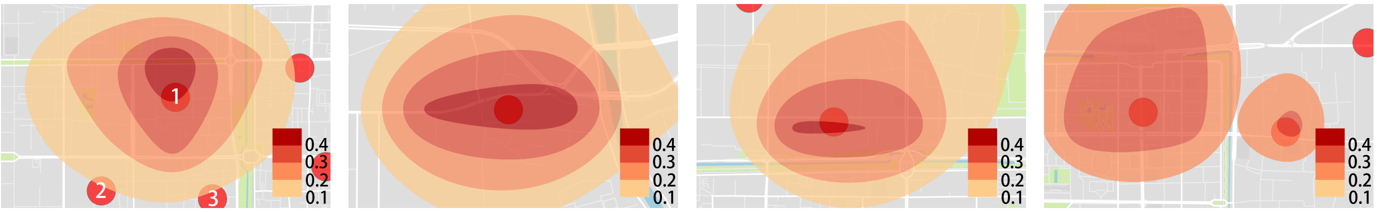
In this series of experiments, we identify the key metrics of attractiveness and assess the accuracy of the Huff model fitted by CellTradeMap to predict the trade areas of other commercial districts using 5-fold cross validation. Specifically, the commercial districts are divided randomly and evenly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

1) Sensitivity Analysis of Attractiveness Metrics: In this experiment, the sensitivity parameters $\gamma_1, \gamma_2, ..., \gamma_H$ are solved from Eq.(9) and each parameter corresponds to a metric of



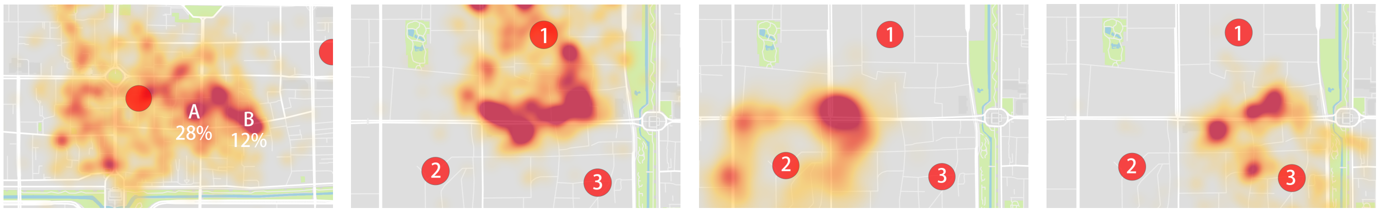
(a) (b) (c)

Fig. 6. (a) Correlation between the number of residents inferred and that by census for each administrative district. (b) The top 8 attractiveness metrics with high sensitivities. The error bar is the variance over 5-fold cross validation. (c) 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) (b) (c) (d)

Fig. 7. Contour maps of visit probabilities. 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. V-C. (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.



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

Fig. 8. Heatmaps showing the distribution of visitors' homes. 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. 7a. (b), (c) and (d) show the distribution of visitors of commercial districts 1, 2, 3 respectively.

attractiveness. The sensitivities are averaged over 5-fold cross validation and the metrics with top sensitivity are shown in Fig. 6b. The results reveal that abundant catering options and parking lots, easy access to public transportation, as well as large numbers of passengers are critical to the attractiveness of a commercial district.

The wide spatial coverage of MFR is crucial for valid sensitivity analysis. Fig. 6c plots the results of the sensitivity analysis with only 5 randomly selected commercial districts. Compared with Fig. 6b, the error bars (the variances over 5-fold cross validation) are much larger in Fig. 6c. This indicates that sensitivity analysis of attractiveness metrics using data of a small number of commercial districts tends to be instable, which is the case in previous studies [9], [10].

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

$$RMSE = \sqrt{\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(P_{ij} - \hat{P_{ij}})^2} \qquad (11)$$

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.

TABLE II
AVERAGE RMSE ON PREDICTION ACCURACY.

Method	RMSE
Linear Regression	0.188
Random_5	0.145
Random_10	0.154
$L^1 - norm$	0.128

- *Random.* Linear Regression with 5 or 10 randomly selected metrics to calibrate the Huff model.

Table II summarizes the results from 5-fold cross validation. The model fitted by CellTradeMap yields the best RMSE. *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.

VIII. CONCLUSION

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.

ACKNOWLEDGMENT

This work is supported in part by the National Key Research Plan under grant No. 2016YFC0700100, NSFC under grant 61832010, 61332004, 61572366.

REFERENCES

[1] Z. D. Boren, "There are officially more mobile devices than people in the world," *The Independent*, vol. 7, 2014.

[2] Cisco, "Global mobile data traffic forecast update, 2016-2021 white paper," 2017.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[4] Y. Zhang, "User mobility from the view of cellular data networks," in *INFOCOM*. IEEE, 2014, pp. 1348–1356.

[5] S. Dhar and U. Varshney, "Challenges and business models for mobile location-based services and advertising," *Communications of the ACM*, vol. 54, no. 5, pp. 121–128, 2011.

[6] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, "Geo-spotting: mining online location-based services for optimal retail store placement," in *KDD*. ACM, 2013, pp. 793–801.

[7] S. Yang, M. Wang, W. Wang, Y. Sun, J. Gao, W. Zhang, and J. Zhang, "Predicting commercial activeness over urban big data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 119, 2017.

[8] D. L. Huff, "A probabilistic analysis of shopping center trade areas," *Land Economics*, vol. 39, no. 1, pp. 81–90, 1963.

[9] Y. Wang, W. Jiang, S. Liu, X. Ye, and T. Wang, "Evaluating trade areas using social media data with a calibrated huff model," *ISPRS International Journal of Geo-Information*, vol. 5, no. 7, p. 112, 2016.

[10] Y. Qu and J. Zhang, "Trade area analysis using user generated mobile location data," in *WWW*. ACM, 2013, pp. 1053–1064.

[11] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, "I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application," in *CHI*. ACM, 2011, pp. 2409–2418.

[12] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2016.

[13] W. J. Reilly, *The law of retail gravitation*. WJ Reilly, 1931.

[14] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: a survey of research," *ACM Computing Surveys*, vol. 47, no. 2, p. 25, 2015.

[15] X. Zhang, Z. Yang, W. Sun, Y. Liu, S. Tang, K. Xing, and X. Mao, "Incentives for mobile crowd sensing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 54–67, 2016.

[16] L. Ferrari, M. Mamei, and M. Colonna, "People get together on special events: Discovering happenings in the city via cell network analysis," in *PerCom Workshops*. IEEE, 2012, pp. 223–228.

[17] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, and C. Peng, "Spatio-temporal analysis and prediction of cellular traffic in metropolis," *IEEE Transactions on Mobile Computing*, 2018.

[18] E. Thuillier, L. Moalic, S. Lamrous, and A. Caminada, "Clustering weekly patterns of human mobility through mobile phone data," *IEEE Transactions on Mobile Computing*, vol. 17, no. 4, pp. 817–830, 2018.

[19] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, p. 779, 2008.

[20] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *MobiCom*. ACM, 2014, pp. 201–212.

[21] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," in *IMC*. ACM, 2015, pp. 225–238.

[22] T. Liu, Z. Yang, Y. Zhao, C. Wu, Z. Zhou, and Y. Liu, "Temporal understanding of human mobility: A multi-time scale analysis," *PLOS ONE*, vol. 13, no. 11, p. e0207697, 2018.

[23] W. J. Reilly, "Method for the study of retail relationships," *University of Texas Bulletin*, vol. 2944, 1929.

[24] E. Dramowicz, "Retail trade area analysis using the huff model," *Directions Magazine*, vol. 2, 2005.

[25] R. A. Peterson, "Trade area analysis using trend surface mapping," *Journal of Marketing Research*, vol. 11, no. 3, pp. 338–342, 1974.

[26] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with gps history data," in *WWW*. ACM, 2010, pp. 1029–1038.

[27] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *WWW*. ACM, 2009, pp. 791–800.

[28] L. Qi, Y. Qiao, F. B. Abdesslem, Z. Ma, and J. Yang, "Oscillation resolution for massive cell phone traffic data," in *MOBIDATA*. ACM, 2016, pp. 25–30.

[29] Y. Jin, N. Duffield, A. Gerber, P. Haffner, W.-L. Hsu, G. Jacobson, S. Sen, S. Venkataraman, and Z.-L. Zhang, "Characterizing data usage patterns in a large cellular network," in *CellNet*. ACM, 2012, pp. 7–12.

[30] P. Yang, T. Zhu, X. Wan, and X. Wang, "Identifying significant places using multi-day call detail records," in *ICTAI*. IEEE, 2014, pp. 360–366.

[31] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[32] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Pervasive*. Springer, 2011, pp. 133–151.

[33] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *KDD*. ACM, 1996, pp. 226–231.