Trade Area Analysis using User Generated Mobile Location Data

Yan Qu
PlaceNous.com
yan.qu@acm.org

Jun Zhang
Pitney Bowes Inc.
jun.zhang@pb.com

ABSTRACT

In this paper, we illustrate how User Generated Mobile Location Data (UGMLD) like Foursquare check-ins can be used in Trade Area Analysis (TAA) by introducing a new framework and corresponding analytic methods. Three key processes were created: identifying the activity center of a mobile user, profiling users based on their location history, and modeling users' preference probability. Extensions to traditional TAA are introduced, including customer-centric distance decay analysis and check-in sequence analysis. Adopting the rich content and context of UGMLD, these methods introduce new dimensions to modeling and delineating trade areas. Analyzing customers' visits to a business in the context of their daily life sheds new light on the nature and performance of the venue. This work has important business implications in the field of mobile computing.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems

General Terms

Human Factors, Measurement

Keywords

Trade Area Analysis, Mobile marketing, Location based marketing

1. INTRODUCTION

One of the defining characteristics of mobile computing is user mobility. Various applications and services have been developed to make use of location information from people's smart phones, ranging from simple apps like WeatherChannel, to location based guides like Yelp, context aware assistants like GoogleNow, and location based social networking services like Foursquare. These new apps and services were widely adopted and as a result, location data from millions of users has been captured. This data, although anonymized, records users' locations and movements over time. Thus, it has great potential in the study of location related behavior.

Significant research has been done using location data of mobile users. Some were fundamental research such as Song et al.'s work on identifying patterns of human mobility [31]. Some focused on building new services that may have great public or business potentials [35], such as modeling city living neighborhood [6] and recommending friends and locations [38]. The business value of the location data of mobile users has generated a lot of attention in research and industry circles. For example, Baccelli and Bolot [1] modeled the economic value of this data. Companies have started mining such data for location based marketing purposes, such as user profiling (e.g. SenseNetworks.com) and location based advertising (e.g. placecast.com).

This work explores a new business application area for location data of mobile users: studying the interactions between customers and local stores. Specifically, we looked to answer critical questions in business marketing management: What is the trade area of a business? Where are its customers from? What are the characteristics of these customers? Answers to these questions are vital for understanding consumer behavior and effectively allocating limited resources [27].

Our work builds on the traditional Trade Area Analysis (TAA) model that has been used in industry for more than 90 years [15][28]. We propose a new framework and corresponding analytic methods to accommodate mobile location data which has a very different nature from data used in traditional TAA. For instance, location data of mobile users usually does not contain users' home address information, which is required in the traditional TAA. At the same time, mobile location data adds new dimensions to TAA process as it has much richer content and context of people's location history. Using a mobile location dataset of Foursquare check-ins, we illustrate the value and challenges of our TAA process in the era of mobile computing.

In summary, this paper makes three primary contributions:

- To the best of our knowledge, it is the first work to apply Trade Area Analysis to location data of mobile users, opening a new business application area.
- It provides a new analytic framework based on the traditional TAA model. This paper focuses on demonstrating the process of applying this framework to mobile location data rather than analyzing the results gleaned from a special dataset (i.e. Foursquare check-ins).
- It presents new analytic methods within the TAA framework to model customer mobility, create customer profiles and preferences, and examine interactions between customers and stores.

The paper is organized as follows: First, we describe the nature and limitations of location data of mobile users, as well as the Foursquare check-in dataset used in this paper; second, we describe the TAA processes using this new data; third, we discuss novel analyses beyond the traditional TAA process; and last, we discuss related works, review the implications of our findings, and make overall conclusions.

2. USER GENERATED MOBILE LOCATION DATA (UGMLD) AND ITS LIMITATIONS

2.1 Location Data of Mobile Users

There are different types of location data of mobile users with varying methods of collection and degree of granularities. We first exclude several types of data that were not applicable to our proposed analyses so they will not confuse readers. The most widely used mobile location data in academic research is Call Detail Records (CDR) [31]. It records the closest cell phone tower location when the phone is used and can only be collected by mobile service carriers. Mobility research often uses CDR because it has low user bias and better time coverage [9][31]. However, the location precision of CDR is about 3 square kilometers on average – the range of cell phone tower coverage [9]. Thus, CDR can only draw conclusions at the area level. Similarly, IP address based location data is collected by websites by checking the location associated with the IP address of its visitors. Neither of these two types of data provide the location precision required for our analyses and therefore they were excluded from discussion. Another type of mobile location data we excluded is collected by in-store sensors (e.g. Bluetooth, WIFI, or ultra sound sensing). Such data is limited to a store's boundary and thus inapplicable in our analyses.

Two types of location data of mobile users can be used in our analyses: the first collected by social networking services (e.g. Foursquare) [16] and the second collected by various apps that record users' locations with their permission (Figure 1.a). Such location data has high precision, usually at the store level. Because they need users' active participation or permission to be collected, we called them User Generated Mobile Location Data (UGMLD).

The UGMLD from location based social networking services is well-known in academic circles because it is easy for researchers to access. For instance, [4] reports that about 20% of Foursquare users automatically publish their Foursquare check-ins on Twitter. Several research groups have used this data source to study location behaviors and population dynamics [4][5][13][22]. Their works indicate that although self-selected user location data is limited and biased, it can be used to identify valuable patterns about individuals and places.

The UGMLD collected by various mobile apps is less well-known in academia but widely used in industry by app developers and third parties like mobile ad networks. For instance, Macy's can collect users' location history in its mobile app if the location feature is enabled (Figure 1.a). Mobile ad networks merge user location data from multiple apps into a more complete location history and then display a location related banner in the app (e.g. Figure 1.b: a banner showing the location of a nearby Wells Fargo Bank.). Note that these apps only record latitude/longitude information. High precision reverse geocoding technology [36] is needed to find venue information.

(a) Macy's (b) WeatherChannel

Figure 1. Two mobile app examples

However, UGMLD has its limitations, including sparse time coverage of location histories and the likelihood to be biased. When a person uses a mobile application, a piece of data is created with the user's current location and a timestamp. This data is generated at a rate ranging from several times a day to once in several months. Therefore, only small fractions of a user's location history are represented, not to mention the complete reconstruction of the trajectories. There are some applications (e.g. Aloha.com) that run in the background and track a user's complete location trajectory, but they have not been widely adopted.

The nature and frequency of app usage depends on its type and user preferences. For instance, searching for a hotel on Priceline happens less frequently than checking in on a social network service like Foursquare. Even for the same app, a person who wants to share his locations with friends will use it differently than one looking for check-in based discounts [33]. Therefore, UGMLD cannot be treated as a simple location history. We must carefully examine how incentives, users' preferences and habits, and contexts will affect the dataset.

Lastly, location privacy is becoming an important topic in research, industry, and government policy making circles [14][18]. In this paper, we used the publicly available datasets and focused on the technical process instead of privacy policies. However, we hope our work will help others become more aware of the related privacy implications of location information.

2.2 Foursquare Check-in as UGMLD

In this paper, we use a set of publically available UGMLD from Foursquare to conduct our Trade Area Analysis (TAA). Note these analyses can also be applied to mobile app data. In this section, we describe the data set and discuss the challenges of using it in TAA.

2.2.1 Our Dataset

Check-ins from Foursquare is frequently used UGMLD by researchers because of its relatively large size and easy access. Our dataset contains automatically published check-in tweets on Twitter from linked Foursquare accounts. It covers a ten month period in 2012 from January 1st to November 1st. It was collected using a similar method described in [4]. There are total 31,554,516 check-ins at 980,686 distinct places from 1,016,181 unique users. The data includes latitude, longitude, time of the check-in, place name, place category, and related information like tips and shouts.

Although the dataset appears large, it is only a small portion of the total data on Foursquare. For instance, while the Foursquare page of the Whole Foods store at Union Square, New York City shows 16,678 unique visitors, our dataset only contains 682 (about 4%) of them. Still, this number is larger than the number of customer participants in many traditional TAA surveys. It is important to note that, people who publish their check-ins on Twitter might be less privacy sensitive than general users and this fact is likely to bias our results. That said, this is the best dataset we can find to illustrate the potential of using UGMLD for Trade Area Analysis.

Aware of the limitations of our dataset, we are very cautious when interpreting results, focusing instead on illustrating novel methods and new directions for further research. Our methods should be more reliable and applicable to a variety of businesses if conducted by those who have access to larger UGMLD datasets, such as Foursquare or mobile ad networks.

2.2.2 Uneven Business Category Distribution

In our dataset, the numbers of check-in by business category are unbalanced. Figure 2 shows the top 80 frequently visited categories from our data set. "Grocery or Supermarket", "Restaurant", and "Coffee Shop" were very popular, while others like "Doctor's office" and "Bank" were much less common. There are several reasons that might explain this skewed distribution. Firstly, as Lindqvist et al. [17] found in their study, people have different motivations for checking in or not based on the location in question. For instance, they may check-in at a coffee shop to "signal availability" or "coordinate with friends", or at places like a gym as "a form of presentation of self". Lindqvist et al. found that Foursquare users do not want to check in at Fast food places because "It's embarrassing to be seen there". Secondly, Foursquare users are technology savvy, and are more likely go to technology-oriented locations than the general public. Finally, the adoption of Foursquare is geographically uneven with most users living in big cities like New York City. Therefore, the places they visit are more common in big cities.

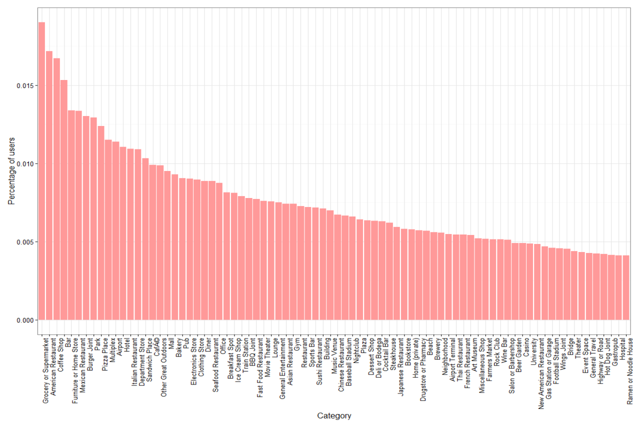


Figure 2. Top 80 check-in categories

Given the skewed distribution by category, the Foursquare data will be more reliable for TAA in the popular categories, not only because there is more data, but also because self-selected data is less biased when the population is larger.

Note: Foursquare also supports none-place check-ins to events. For instance, the most popular "place" in New York in 2011 was a weather event "Snowpocalypse" [32] that could be checked in from any location. These kinds of check-ins were excluded during our collection process.

2.2.3 Skewed User Check-in Distribution

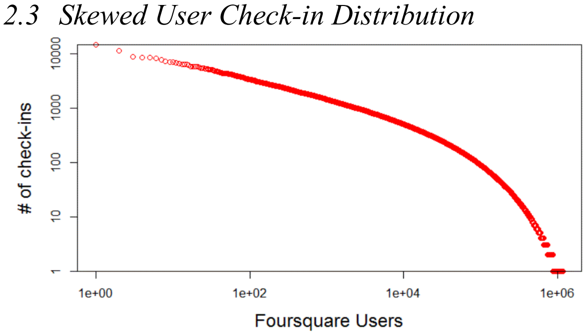


Figure 3. Check-in distribution of Foursquare users

Not surprisingly, the check-in frequency per user is also highly skewed (Figure 3). The number of check-ins range from 1 to 14406 with the median being 5. 25.6% of the users only had one check-in. They might have only tried Foursquare once or they stopped sharing their check-ins on Twitter. There were two users with more than 10,000 check-ins. A detailed review revealed that both were software robots that publish local news. These outliers indicate that we should be very cautious when using check-in data to infer user behavior. In this study, we interpret the results carefully, well aware of potential biases introduced by these skewed distributions.

3. TRADE AREA ANALYSIS USING FOURSQUARE DATA

3.1 General TAA Process

Trade Area analysis gives a business information about where its customers are from, how far they travel to the store, and their demographics and household information. Thus, marketing activities such as direct mail campaigns can be tailored accordingly.

General Trade Area Analysis typically contains the following steps:

1. Collect basic information about the store to be modeled, such as location and store type, both of which have a large impact on the store's trade area.
2. Select a sample of current customers and collect related customer information, especially where they are from (usually their home address) and their spending in the store.
3. Derive the travel distance polygon to identify the geographic boundary of the customers' locations.
4. Identify the user block group and related information. Create customer profiles, such as block group "Psyte" profiles [19] that represent people's demographic information and income levels, so the business can better understand its customer type.
5. Incorporate other factors like competitors and sister stores.

In following sections, we explore whether we can generate trade areas using Foursquare check-in data, and discuss the challenges of using this type of data. Our analysis roughly follows the same steps of the general TAA process.

3.2 Our Store Samples

The trade area of a store is shaped by many factors, including but not limited to business type, customer population (e.g. students, local residents, daytime employees), geographic settlement context (e.g. urban vs. suburb), underlying road network, and competition. For instance, in traditional TAA, convenience stores that provide products needed on a regular basis (e.g. grocery stores and coffee shops) are treated differently than destination stores that provide major products such as furniture or appliances.

Table 1: selected stores

Store Name	Category	Location	# Checkin Customers
Whole Foods	Grocery store	Union Sq., NYC	682
IKEA	Furniture store	Canton, MI	380
Starbucks	Coffee shop	Union Sq. NYC	420
Macy's	Department store	Downtown, SF	120

To incorporate a variety of the major factors in TAA, we choose four stores of different types and locations as samples, shown in table 1. Most of our analyses that involve store comparison use the Whole Foods in NYC and the IKEA in Canton, MI because they differ in store type and location demographic: the first is a grocery store in an dense urban setting serving residents and daytime employees; the second store is a furniture store in a suburban area that mostly serves residential customers. We also used the other two stores to highlight some interesting observations. Please note: we limited our sample selection to stores with at least 100 unique users in our dataset (a traditional

TAA practice). Thus, our sample is biased towards popular stores in big cities.

3.3 Check-ins as Customer Visitation Data

In traditional TAA, businesses gather customer data from transaction records of current customers, membership information, and surveys of nearby areas. All these methods have their limitations. It is difficult for a business to get its customers' home or office address unless it is a delivery business or requires a registered membership. Business usually only get zip code level information from credit card transactions due to government regulations. Surveys or focus groups are regularly used but expensive and time consuming. They also have issues with self-reporting bias. Although the check-in data differs from actual customer visits, it is relatively easy to get and offers valuable insights into customer behavior, particularly for businesses who want to reach their customers or potential customers with mobile or online marketing.

Below are examples of the check-in data. Note that the latitude/longitude data we retrieved from Twitter is Foursquare normalized store location instead of the real GPS location sent to Foursquare originally.

```
[user_id], [checkin_time], [latitude], [longitude], [store_id]
196514, 2010-07-24 13:45:06, 53.364811914, -2.2723465833, store21
245677, 2010-07-24 13:44:58, 53.364811914, -2.2723465833, store21
....
196514, 2010-07-25 11:21:43, 53.364811914, -2.2723465833, store21
```

In our data, some users checked in at the same store multiple times while many only checked in once. Figure 4 shows the uneven check-in distribution of the four stores with varying degrees of skewness. The top 20% of Starbucks' customers made about 70% of Starbucks' check-ins, while the top 20% of IEKA's customers made only about 41% of IKEA's check-ins, indicating that IKEA has a relatively even user check-in distribution.

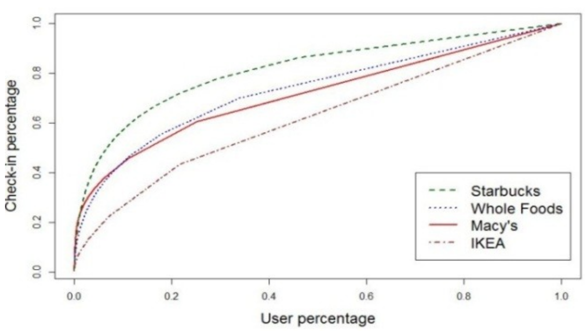


Figure 4. Store check-in distribution

Although there are Foursquare-specific reasons that can lead to this uneven distribution, such as small number of people aggressively checking in to fight for the mayor-ship of a store, these results still reflect customer-store visiting patterns. Stores have loyal customers who visit the store regularly and random customers who rarely visit. Starbucks is a convenience venue thus it is more likely to have customers who visit everyday, while IKEA is a destination venue and therefore likely to have fewer daily visitors.

Furthermore, considering people's check-in behaviors, there are much less false positives (a person checked-in but did not visit) than false negatives (a person visited but did not check in). Therefore, frequent check-ins are likely to indicate a frequent customer, but frequent customers may not check-in frequently. Thus, the check-in distribution is the lower bound of the actual visit distribution. However, we cannot tell if the actual visit distribution is flatter or more skewed.

The important question is how the skewness in data will affect the Trade Area Analysis. For some analysis, the skewness will not bias the results. For example, when drawing trade area polygons, we estimate one activity center for each sampled customer, thus the process is indifferent to the skewness of personal check-ins. However, the skewness will bias the results for certain analysis. For example, in sequence analysis, we want to know where people usually visit before their visit to a specific venue. If the data is highly skewed toward a small group of frequent customers, their behavior pattern will dominate the results. Therefore, in our analysis, we identify such frequent customers and sample only a small number of sequences from them. Throughout this paper, we will adjust the data according to the nature of the analysis and explain our rational.

In our work, to counter the impact of over aggressive check-ins by some users, we removed duplicated check-ins at a venue at the same day. That is, for one customer, we considered at most 1 check-in at one venue in one day. We call such check-in the venue-day check-in (VD check-ins). This is a more truthful indicator of how important a venue is in the person's daily life. The following analyses were only made on VD check-ins.

3.4 Derive Trade Area Boundary using Check-ins

One of the basic questions TAA tries to answer is "Where do my customers come from?" Traditionally, home addresses are collected from a sample of existing customers and used to generate a polygon that includes the majority of customer's households.

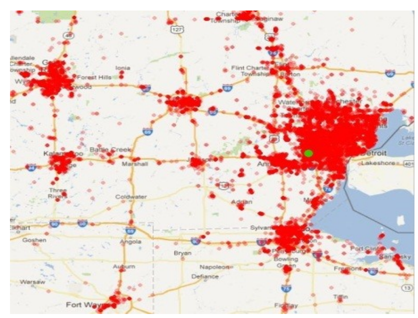


Figure 5. Check-in locations of IKEA customers

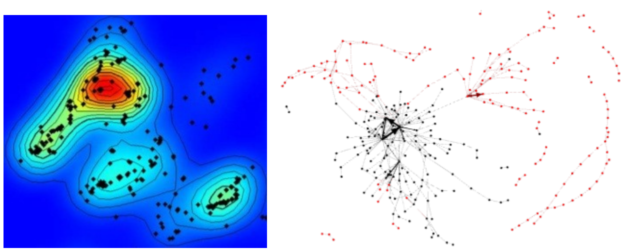
However, there is no explicit home address information in our check-in data. Instead, we have a list of customers' check-in locations. Figure 5 shows the spatial distribution of check-in places from customers of the IKEA store at Canton, MI. The picture was generated following two steps: we first identified all the customers who checked in at the IKEA store (the green dot). Then we identified all other places these users checked in and posit them on the map (red dots). Instead of explicitly telling us where the customers are from, this map tells us the areas the customers frequent. According to the basic geographic law of distance decay: interaction between two locales decreases as the distance between them increases, most check-ins a person made should be close to important places in his life, such as his home or office [12]. Therefore, this map provides a rough image of where these important places may be located for this group of IKEA customers.

However, because of the high skewness of the check-in distribution, the patterns shown in Figure 5 may be biased toward those customers with many check-ins. To present the trade area more faithfully, we need to estimate the important places for each customer.

3.4.1 Identifying Users' Activity Centers

In our approach, instead of identifying a customer's home, we identify important places or areas in a customer's life using his check-in location histories.

Previous research based on CDR and UGMLD [4][9][12][25][31] indicates that people's check-ins and activities exhibit both place and area regularity. First, there are places people regularly frequent [9][12][37]: besides their home and office, this might be a gym, grocery store, or library. Second, there are regularly visited areas even if no specific store or location in that area is regularly visited [25]. As an example, figure 6a shows the check-in location heat map of a person who voluntarily shared his Foursquare check-ins and related location information. From the map, we can see that the majority of check-ins falls in three clustered areas. The largest one on the top is the person's "work cluster". The hottest zone with red color in this cluster is around the person's office. The stripe shape of this large cluster is formed by a large number of check-ins along a main street close to his office where he usually has lunch. The bottom right cluster is around the person's home. The cluster in the middle is a popular shopping area. Figure 6b shows the trajectory network of the same user's check-in sequences. It is generated by linking any two consecutive check-in places that happened in the same day. From this figure, we can see there are several check-in hubs with high network degrees. Naturally, the hubs are located in the clustered area in the heat map, which means that frequently visited areas have both spatial importance and sequential significance.



(a) (b)

Figure 6. Check-in heat map and trajectory network from a same person

Place and area regularities allow us to extract locations where there is a high probability that a user will show up in the surrounding area. We call these locations activity centers.

For the purpose of TAA, activity centers are a valid substitute for home location for the following reasons. First, home locations are used to identify target marketing areas. With mobile marketing technologies, knowing where a customer is and where he is likely to show up is more important than only knowing where to mail advertisements. Activity centers serve this purpose better than traditional home location. It is important to note that a person may have multiple activity centers (e.g. home and office). Businesses have a better chance of attracting a customer when they are located close to any one of the customer's activity centers. Second, home locations are used to determine community level demographics or economic data to create customer profiles in traditional TAA. Extracting multiple activity centers, we have a good chance of locating one close to the customer's home (see below for details). Therefore, we can still make coarse community level inferences. Moreover, we propose new methods of creating customer profiles using UGMLD (section 3.5), which generate useful customer information that was never available before.

Previous works such as Isaacman et al. [12] indicate that it is possible to identify important places like home or office area using CDRs. Building on previous works, we propose four methods to identify activity centers using UGMLD and then conduct a limited test.

- Center of Mass: It locates the spatial center of all check-in locations of a person, weighted by the number of check-ins in each place. Note that this method may result in a location in between several check-in clusters.
- Most frequently checked-in location: Intuitively, a customer may be more likely to show up in an area close to this location than other areas.
- Location with the highest check-in density. This method tries to find the hottest spot in a check-in heat map.
- Center of mass of the most frequently visited location cluster. This method firstly identifies a person's regularly visited areas - clusters in the heat map, then extracts the mass center of the most frequently visited cluster. We used the popular density based clustering algorithm DBSCAN [8] with certain prefix density threshold (i.e. there must be at least 5 check-ins in a 1km radius over a period of 30 days to become a potential cluster center). More important is an extension of this method: selecting mass centers of multiple most active areas and then choosing the activity center from them accordingly in TAA. This method heeds the fact that people usually have multiple regular activities areas (e.g. home area and office area). Different TAA tasks may use different activity areas. For example, we can pick the closest activity area rather than the home location when doing a mobile targeting analysis (e.g. for stores whose customers are mainly office workers).

We conducted a simple test to explore the relationship among the activity centers extracted by the above methods and the actual home location. We identified all users who have explicitly checked in at home and have more than 200 check-ins during the 10 month period in our data set, which resulted in 466 users. Figure 7 shows the distance from the identified activity centers and a person's home. The method "cluster – 1" selects the mass center of the most active location cluster. The method "cluster – 3" selects the mass center closest to home from the top 3 active location clusters. Clearly, when we take the top 3 most active clusters, it is very likely that one of them is close to the home location. In our test, 293 (64%) out of the 466 persons live within 2 miles one of the cluster centers. We use this method in the rest of paper.

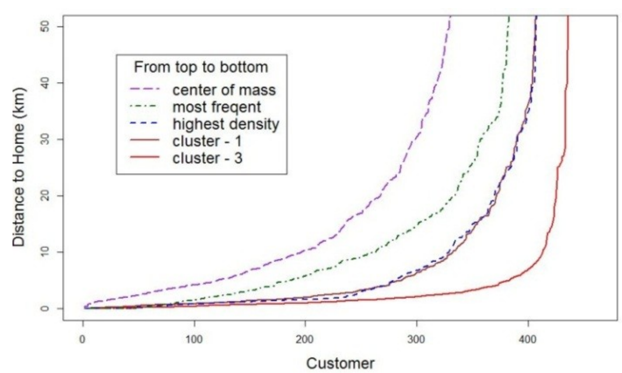


Figure 7. Performance of different activity center identification algorithms

We are aware that this analysis is heavily biased toward people who shared their home locations and these people may have different check-in patterns than others. However, the focus of this paper is not to develop a home prediction algorithm, but rather to demonstrate the added benefits of using activity centers in TAA. Although highly correlated to home locations, activity centers have their own characteristics and meaning within TAA. Many proposed analyses in this paper can be applied to both home location and activity centers. Again, we must emphasize that the results are only valid and valuable if the analysts know how to correctly interpret the data.

3.4.2 Drive Distance based Trade Area Boundary

There are several ways to derive the trade area boundary of a store. Common methods include drive-time/distance polygon, data driven rings, and density clusters [15]. In this paper, we used the drive-time/distance polygon because it is more accurate than the data driven rings and it is less data demanding than the density cluster method which requires large number of user samples to stable the clustering results.

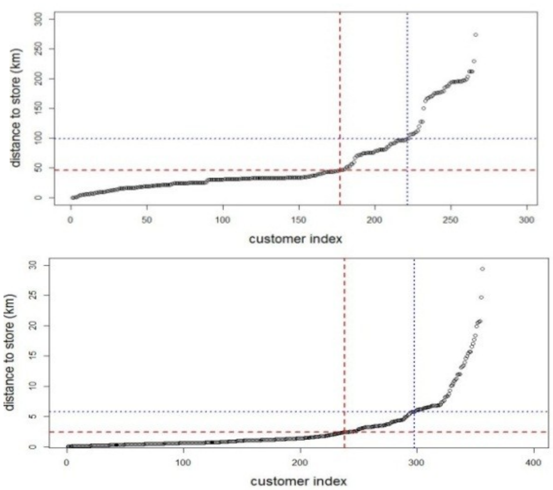


Figure 8. The distance decay of the IKEA (top) and Whole Foods (bottom)

The construction of drive-time/distance polygons starts with distance decay analysis. Distance decay is a geographical term that refers to the decrease of cultural or social interactions as distance increases. In TAA, the most obvious distance decay is the decrease of the number of customers' households or activities centers as the distance to the store increases. Figure 8 illustrates the distance decay effect using the IKEA and Whole Foods stores as examples. The x-axis is the customer index. The y-axis is the distance between the store and the closest activity center of each user. Traditionally, the primary trade area is defined as containing 60% or 75% of the customer population. In our examples, about 60% of the MI IKEA's customers are within 50KM (red line) distance and 75% customers are within 100KM (black line) distance. The Whole Foods at NYC has a much faster distance decay, thus it has a much smaller trade area. 60% of its customers are within 2.5KM and 75% are within 5.8KM.

Figure 9 shows the visualization of the primary trade area polygons containing 60% and 75% customer population of the two stores. The red dots are identified activity centers of customers who have checked in at the store. In the IKEA map, we used the drive-time polygon because people's movement is highly influenced by the road network like high-ways in the suburbs. In the NYC Whole Foods map, we used the drive-distance polygon because it is short ranged and the road speed differences are unpredictable.

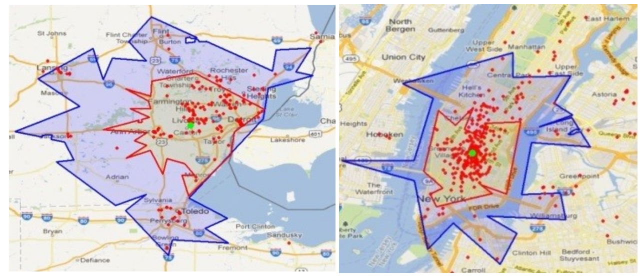


Figure 9. Drive distance/time polygons of two stores (left: IKEA, right: Whole Foods)

These two maps and related analyses reveal a lot of information to businesses and can help them with location related decisions. For instance, these stores now know where their customers are from and how long it takes for then to get to the store. The MI IKEA store may want to purchase mobile ad banner space for people who are active in their trade areas such as "Livonia, Detroit, Ann Arbor" while the NYC Whole Food may focus their marketing at lower and midtown Manhattan.

Furthermore, although we did not directly compare the drive-time/distance polygons generated using check-in data to the ones generated using real customer visits, we talked with several trade area experts in industry and their feedback was that the trade areas we created for these stores made great sense and the drive distance patterns were similar to those found using traditional TAA at similar stores.

The trade areas generated here might be biased due to the self-selective nature of UGMLD and our limited data sample. Nonetheless, this method is still very valuable while other data is hard to get. The bias will be less of an issue if analysis is limited to populations similar to the sampled customers or when researchers have access to more robust UGMLD.

3.5 Location-based User Profiling

Generating the drive-time/distance polygon is not the final step of TAA. In practice, businesses want to know more about current and prospective customers such as their shopping habits and demographic background. Transaction histories and membership information collected by stores are often used to profile customers. However, such detailed personal information is not always available. A less satisfactory approach is Geo-profiling [15][19] – a commonly used method to approximate user characteristics based on neighborhood demographic data (within a 0.3-2 miles range). Its precision is limited but it is very useful when better information is unavailable. However, this method may not be applicable to our dataset, which contains no home address information (activity centers may be too far away to generate similar demographic data.).

To demonstrate new possibilities of user profiling using UGMLD, we choose to characterize users' check-in patterns using the place category information. This new profile feature will tell us what types of places a person like to visit, which is valuable for mobile marketing purpose.

An easy solution is simply count the frequency of one's check-ins within each category. However, we found that almost all users checked in under popular categories like coffee shops. Therefore, this method cannot efficiently distinguish different check-in patterns. Moreover, Foursquare has a hierarchical category structure which includes 9 top categories and 410 sub-categories.

Sub-categories are more valuable in user profiling as they separate a person who likes "Italian food" from another one who likes "Korean food". However, the high dimension of the 410 subcategories presents a challenge.

To address these issues, we chose the Latent Dirichlet Allocation (LDA) [3] method to identify hidden check-in patterns and to profile users based on the identified patterns. LDA is widely adopted in document topic modeling. It assumes that each document contains a mixture of topics and each topic has certain probability of mentioning a word. LDA identifies topics and calculates the proportion of different topics in each document by examining word distributions in the documents. In our work, we treat each person as a "document", each place category as a "word". The "topics" are hidden check-in patterns in the population. LDA requires a pre-determined number of topics. After testing different numbers, we choose 6 that gives intuitively meaningful results.

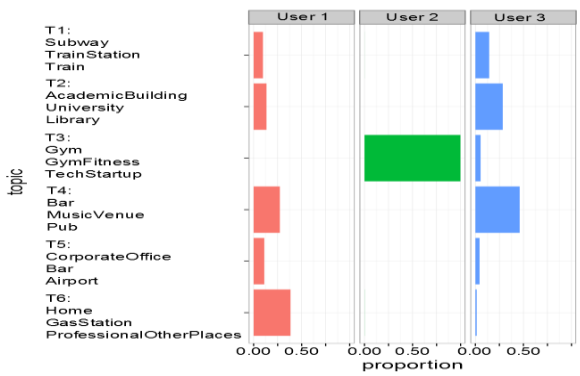


Figure 10. LDA "topics" and user profile

Figure 10 shows the 6 topics (T1 – T6) identified in our dataset, reflecting different check-in patterns. Each topic is labeled by three location category words that have the highest latent factor weight for that topic. T1 is daily commute; T2 is life at school; T3 is the gym and fitness; T4 is night-life; T5 is corporate life and travel; and T6 is the home, driving, and work. We then profile users using these topics. Figure 10 also displays three sample user profiles. We can easily distinguish User 2 from User 1 and User 3 as he only checked in under T3. From his profile, we may guess that he had a gym membership or might work at a startup. The difference between User 1 and User 3 is more subtle. Both had an active night life and traveled to locations on a campus, but User 3 is more likely to be a student in the city than User 1 because he was more active in T2 but almost had no activities in T6. Note that popular categories such as coffee shops may not show up in the LDA results, because they are not helpful when trying to distinguish different groups of users. LDA also addresses the high dimension problem by grouping subcategories into topics.

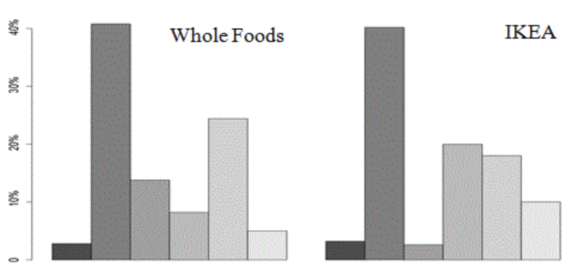


Figure11. User profile comparison

Next, we compare users' profiles for different stores. To make the comparison easier, we profile each user by the topic with the highest score. Figure 11 shows the profile distributions for Whole Foods and IKEA. Both stores have a significant portion of T2 users (probably college students). The biggest point of difference is at T4 (night life goers).

Our profiling method can be further improved. A more sophisticated method is described in [13]. Our limited goal in this section is to show that the location history based profiling is a promising alternative to traditional profiling methods in TAA.

3.6 Competition and Gravity Models

Competitor stores have a large impact on Trade Area Analysis. Gravity models such as Reilly's Law of Retail Gravitation [28] and Huff's Law of Shopper Attraction [11] are usually adopted in competitor analysis. They are essentially benefit-cost analyses: how frequently a customer visits a business depends on the benefit received by visiting and the cost to visit that location. However, there is a lack of empirical studies modeling a business's attractiveness and the customer's benefit-cost decision making process, which calls into question the validity of those gravity models. Traditionally, there was almost no way for a business to know whether their customers visit competitor stores except expensive surveys. Fortunately, the rise of UGMLD mitigates this concern by providing needed information at little cost. It opens a new window to model customer's preferences in competitor analysis.

In UGMLD, we have information of many customers' visits to many business venues. Formally, for a customer $C_1 \sim C_n$ and a group of competitor venues $V_1 \sim V_m$, we have

$$\begin{array}{cccc} & C_1 & \cdots & C_n \\ V_1 & a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ V_m & a_{m1} & \cdots & a_{1mn} \end{array}$$

Where a_{ij} is the number of visit of customer C_i to store V_j.

What we want to know in competitor analysis is customer's individual and aggregated preferences among competitor stores:

- $P(V_j)$: among all similar business venues $V_1 \sim V_m$ what is the probability of venue V_j being visited;
- $P(V_j|C_i)$: for a customer C_i, what is the probability of visiting venue V_j among all similar places.

Using UGMLD, we can easily estimate customer's preference:

$$P(V_j|C_i) = \frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}}$$

$$P(V_j) = \frac{\sum_{i=1}^{n} a_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}}$$

Such estimations are not possible in traditional TAA when data is limited to a single store (i.e. one row in the above matrix).

We use Whole Foods as an example to illustrate the estimation of customer preferences under the influence of competitor stores. In practice, the competitor stores are manually picked using knowledge about the store, region, and customer segmentation. We didn't have this knowledge about the NYC Whole Foods store, so we simply used the category information provided by Foursquare and treated all stores in the "Grocery and Supermarket" category as potential competitors. For each Whole Foods customer, we counted their check-ins at the store and total check-ins in the category of "Grocery and Supermarket". We estimated $P(V_j|C_i)$ as the number of Whole Foods' check-ins from customer C_i divided by the number of grocery store or supermarket check-ins from C_i. We call $P(V_j|C_i)$ the loyalty index because it reflects how likely a customer will visit this store among all competitors. Note: the Foursquare data may be biased because Whole Foods is not a regular grocery store.

Figure 12 shows a map of the Whole Foods customers with their loyalty index. In the figure, the black dot in the center is the store location. Each colored dot represents a customer. The location of a colored dot is a customer's activity center (we used 5 check-ins in a 0.5km radius area within 30 days as the density threshold in our clustering algorithm). The size of a dot is in proportion to a customer's total grocery or supermarket check-ins. The color of a dot represents the loyalty index of a customer, from 0 (blue) to 1 (red).

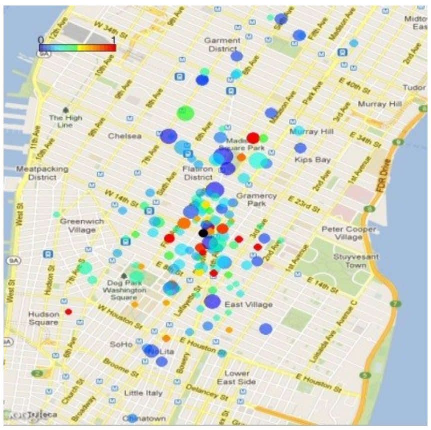


Figure 12. Loyalty index of Whole Foods Customers

This loyalty index will not make sense if a customer seldom visits the store or its competitors. In an extreme case, if a customer has only visited the Whole Foods once and has never visited another grocery store or supermarket, then he will have a very high loyalty index (=1), even if he never returned to the Whole Foods again. Therefore, in practice, we suggest a high threshold for the number of visits. In this paper, in order to present the spectrum of different users, we set a relatively low threshold of 3 visits to "Grocery and Supermarket" stores.

Among the 682 Whole Foods customers, there are 171 who have 3 or more visits to Grocery or Supermarket stores. Their loyalty indexes range from 0.01 to 1, with the media of 0.25. From the map we can see the activity centers of our customers are cluttered around the Whole Foods store. Actually 104 (61%) customers have their activity centers closer than 1 km to the store. All the customers with high loyalty index (>0.5) are within 1.45 km distance to the store. However, our data shows no correlation between the loyalty index and the distance between activity centers and the store. Interestingly, customers with a high loyalty index (red or orange dots) are less frequent grocery shoppers than those who have lower loyalty index (blue dots). One possible reason is that when a customer does a lot of grocery shopping, he is more likely to explore more stores. This map provides a visual representation of the trade area under the influence of competitor stores, giving us a rough idea about how likely customers will visit this Whole Foods store among all of its competitors. With more data, this method will provide a more accurate estimation of shopping probabilities, which can be used to inform business strategies or used as the ground truth for testing competing trade area models.

4. BEYOND TRADITIONAL TRADE AREA ANALYSIS

One of the most promising aspects of UGMLD is that it provides a rich and dynamic context of location-based activities that is not included in traditional TAA. In traditional TAA, the business can only get limited transaction information and some customers' home addresses with related demographic characteristics. UGMLD contains much richer information, such as whether a user visited multiple stores in the same area on the same trip, or where he went before and after the visit to a store, which can be used to infer location trajectories. In this section, we explore how to parse this rich information.

4.1 Distance Decay of Customer's Activity

With UGMLD, we can examine the trade area of a store in the context of customers' daily life. In traditional TAA, we study a store's distance decay and where a customer's household is located within the trade area (e.g. in the center or on the fringe). Taking a customer-centric perspective, we examine the distance decay of each customer's shopping activities and ask the question: where is the store located on its customers' activity areas? This gives us a peek into the role a store plays in its customer's daily life.

For each customer, we calculate the Distance Decay Percentage (DDP): percentage of activities within the r radius area to a customer's activity center, where r is the distance between the customer's activity center and the store location. A high DDP (close to 1) means that the store is on the fringe of the customer's activity area. A low DDP (close to 0) means that the store is close to the center of the customer's activity area.

Figure 13 shows the histogram of DDP from all customers of two different stores: Whole Foods and IKEA. X-axis is the activity DDP. Y-axis is the number of customers falling into each bucket. The difference between the two stores is obvious: for most of its customers, the Whole Foods is close to their activity center. On the contrary, IKEA is located on the activity fringe for a large portion of its customers. This implies that Whole Foods is a neighborhood store with most of its customers living nearby, whereas IKEA is a destination store with many of its customers traveling outside of their major activity area.

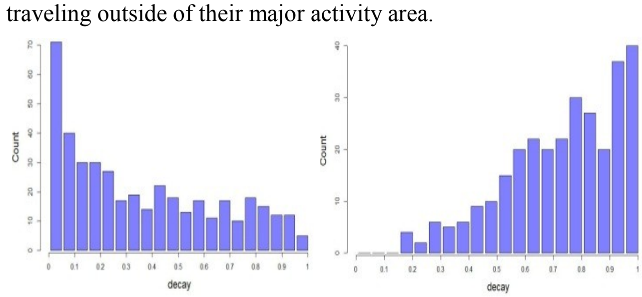


Figure 13. DDP histogram, Whole Foods (left) and Ikea (right)

To further understand the DDP distribution, we draw the scatter plot of DDP vs. distance between store and activity center. In Figure 14, we have plots for Whole Foods and IKEA. Each red dot on the plot is a customer. The size of the dot is in proportion to the log of the total number of check-ins from that customer. The x-axis is the distance between a customer's activity center and the store. The y-axis is the DDP of that customer.

Again, these two plots show two distinct patterns. For Whole Foods, there are a large group of customers with an activity center within 1 km of the store and a relatively low DDP, which means the store is very close to the center of their activity area. When the distance move from about 1km to about 10km, the plot shows a strong correlation between DDP and the distance, which means as customers' activity centers move further away from the store, DDP increases about proportionally. The cluster of dots on the top right corner is customers with activity centers hundreds of miles away from Whole Foods. Understandably their DDP is high, which means the store is on the fringe of their activity area.

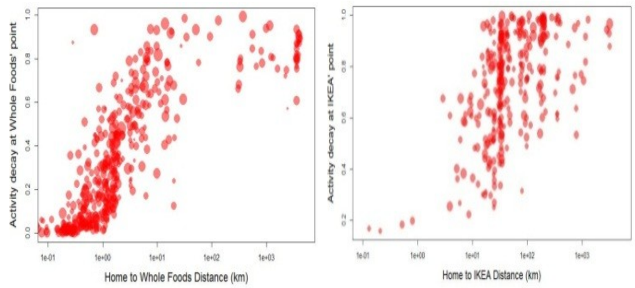


Figure 14. DDP vs. Distance plot, Whole Foods (left) and IKEA (right)

Compared to the plot for Whole Foods, the IKEA has few customers with activity centers within a 5km radius of the store. There is a large group of customers whose activity centers are 10km to 100km away. The almost vertical pattern in the plots indicates a large variety of DDPs at a similar distance, implying a significant difference in customers' mobility (i.e. some customers have a much larger activity area than others). There are also more dots with a high distance but moderately low DDP (0.5~0.8), which are customers who live far away but have very large activity areas, and with the store closer to their activity center than the fringe.

This example analysis shows how a customer-centric perspective can shed light onto hidden patterns of store customer interactions. Above all, we can see that the customer-centric perspective puts the store and purchase transactions in a context of customer's life and helps us understand the intrinsic relationships between customers and stores. This will ultimately help businesses market their products or services and identify potential customers.

4.2 Check-in Sequences Analysis & Use

4.2.1 Analyzing Check-in sequences

Reexamining customers' visits to a particular store in the context of shopping trips can reveal trajectory patterns and provide valuable marketing information. In this section, we explore how to extract and use people's shopping trip information from the check-in data.

To identify shopping trips associated with a particular venue, we extract check-in sequences by sequentially linking the check-ins from a customer during the day he checked in at that business venue.

Next, we analyze the spatial and temporal characteristics of these trips. We use Starbucks to explain what kind of insights can be gained from check-in sequence analysis. We extract 237 sequences with a median number of stops of 3, a median distance span of 2.6km, and a median time span of 5.8 hours.

We draw a bar plot (Figure 15) to illustrate detailed spatial-temporal distribution of check-ins in those sequences. To make the plot readable, we only take check-ins within a 10km distance and ± 10 hour time difference, resulting in 938 (84%) of the Starbucks check-ins in the sequences. In the bar plot, the x-axis is the distance between a check-in in a sequence and the store check-in in the same sequence (before: negative, after: positive), the y axis is the count of check-ins in the corresponding distance range. The colors in the bar indicate the time difference between a check-in and the store check-in, ranging from -10 hours (red) to 10 hours (blue).

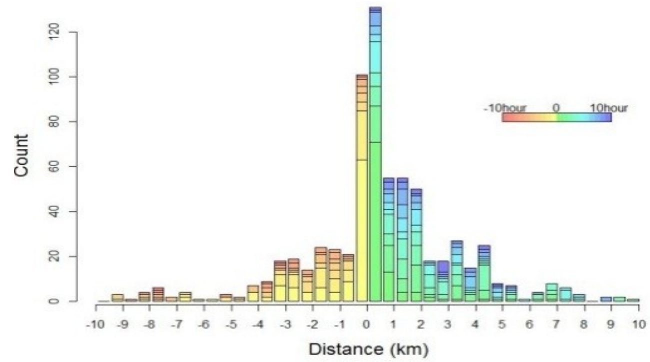


Figure 15. Sequence Check-in distributions (Starbucks)

Several interesting observations can be made from this figure.

First, the highest bars are near to "0" on distance axis within 0.5km, which means many check-ins in the sequence are located very close to the store. Looking at the time segments in those bars, we can see that most of those check-ins are within ±1 hour of the store check-in. This indicates that people check-in at nearby places right before or after they visit a store.

Second, there are more check-ins after the Starbucks check-in than before it, probably because Starbucks is an early destination where people start their trip before moving on to other places. We can infer that people are likely to go to Starbucks to start their day or before other activities.

The check-in sequence analysis provides another example of examining customer-business interactions from a customer-centric perceptive. A business with this knowledge can adjust their marketing strategies accordingly. Data driven geo-fencing is a good example.

4.2.2 Geo-fencing using check-in sequence data

Geo-fencing is a new and exponentially growing mobile marketing technique. The general idea is to target opt-in consumers while they are within a predetermined fence around the store and send them messages or special offers to attract them to walk into the store and make a purchase [20][21]. A key step in the process is setting up the fence. Most current solutions either create a pre-determined radius ring based on store type and location, or use available mall boundaries. The check-in sequence analysis suggests that we can use the check-in data to derive a data driven polygon for Geo-fencing.

Figure 16. Geo-fencing using check-in sequence data

Figure 16 shows a geo-fence created for Whole Foods using check-in data. We used the location a user checked in right before he visited the store. We used the similar distance decay analysis in TAA to determine the drive distance. The red boundary is 0.9km from the store and covers 60% of locations the store's customers visited before they checked in at the store. The blue boundary is 1.8km from the store and covers 75% of such locations. We believe that this fence is more likely to be reliable than a pre-determined radius ring, especially when there is no real consumer feedback data in the early stage of geo-fencing marketing.

Alternative geo-fencing methods using check-in sequence may include using all points in the sequence or clustering the points into zones. But we will leave that discussion for a future work.

5. RELATED WORK

Our work builds on industry practices and previous academic research on Trade Area Analysis. Although 90 years old, the pioneering work by Reilly et al. about Laws of Retail Gravitation model is still used in practice to this day. Other classic early work includes Huff's development of probability based analysis of trade area [11] and similar works such as [7][27][29][30]. Our work also greatly benefited from consulting internal subject matter experts and white papers. That said, we have not found any other publications that apply Trade Area Analysis to mobile location data. We hope this paper will draw more attention to this powerful approach.

Our work is inspired by related work in geo-fencing and mobile marketing [2][20][21][26]. Greenwald et al. [10] discussed geo-fencing solutions that can be deployed to large population and used for mobile proximity marketing or social networking services. Ye et al. [34] demonstrated that it is possible to recommend locations based on people's Foursquare check-in histories. Provost et al. proposed a method of geo-social network targeting for mobile advertising [24]. Many of these findings have already been commercialized. Provost et al.'s work has been used in mobile marketing by EveryScreenMedia.com and similar ideas to Ye et al.'s have been used by SenseNetworks.com. Lastly, Partridge and Begole surveyed existing targeting advertising technologies [23] and described the benefits of activity based marketing. They argued that consumers and advertisers' interests are not necessarily at odds, and a balanced privacy and location context sharing can benefit both parties. We agree with this view and think UGMLD-based TAA fits into this vision.

Our work also benefited from excellent works analyzing and making use of UGMLD and human mobility research. Several of our processes were built on methods developed in these works. The details of these linkages are discussed in the previous sections so we will not cover them here again.

6. DISCUSSIONS

In this paper, we open a new application area for User Generated Mobile Location Data (UGMLD): modeling trade areas and consumer-store interactions. Although the dataset has limitations, we demonstrate that it is possible to build meaningful trade areas based on it, including creating drive distance/time boundaries, generating customer profiles, and weighing competitive factors.

This work has immediate business implications. The most promising application is location based mobile advertising. UGMLD-based TAA can inform businesses about the areas their customers visit. Location histories with rich contextual information can be used to model customer behavior, which is likely to outperform existing geographic block based approaches. Moreover, geo-fences created using the dynamic information in UGMLD will more accurately target potential customers than current geo-fencing practices.

Our work outlines a new framework and corresponding analytic methods for UGMLD based Trade Area Analysis. The unique features of UGMLD makes direct adoption of traditional TAA methods impossible. Three key processes were created as a result of this work: identifying activity centers of users, profiling users based on their location history, and modeling users' preference probability. The identification of activity centers is particularly exciting because it proposes a new location concept that is more suitable to mobile TAA analysis than home location. We also extended traditional TAA by adding new types of analysis such as customer oriented distance decay analysis and check-in sequences analysis. This new customer-centric perspective will provide more insights into the relationship between stores and their customers.

Our paper has several limitations: First, our dataset is limited and could be biased. We discussed its bias in detail as well as the steps we took to minimize it. Second, we limited our trade area analysis to four specific business venues and only provide details of two in this paper due to space limitations. Therefore, caution is needed when interpreting or extrapolating these results. Third, we wished to explore each step of the TAA analysis more deeply, but as the first work covering these new topics, we tried to first and foremost to establish the conceptual framework.

Finally, we wish that we can connect to privacy research more closely in the future, because people's privacy preferences, government regulation, and industry practice directly affect how we tweak and use the processes proposed here. For instance, our process for identifying activity centers could be viewed as intrusive or sensitive to privacy concerns depending how it is applied.

In conclusion, despite the limitations of this paper, we believe this conceptual framework and corresponding analytic methods make important theoretical contributions and provide valuable insights to businesses in fields related to mobile computing.

7. REFERENCES

[1] Baccelli, F., and Bolot, J., "Modeling the Economic Value of the Location Data of Mobile Users." *In 2011 Proceedings of IEEE INFOCOM*, 1467 -1475, 2011.

[2] Benita, R., and Subash, G. V., "Design and Implementation of Location Based Ad Delivery System." *Int. J. on Recent Trends in Engineering & Technology* 5, no. 01 (2011).

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.

[4] Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z., "Exploring Millions of Footprints in Location Sharing Services." In *Proceedings of AAAI ICWSM 2011*.

[5] Chon, Y., Lane, N. D, Li, F., Cha, H. and Zhao, F., "Automatically Characterizing Places with Opportunistic CrowdSensing Using Smartphones." In *Proceeding of UbiComp 2012*. ACM, 2012.

[6] Cranshaw, J., Schwartz, R., Hong, J. I, and Sadeh, N., "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City." In *Proceedings of ICWSM 2012*, 2012.

[7] Donthu, N. "Comparing Market Areas Using Kernel Density Estimation." *Journal of the Academy of Marketing Science* 19, no. 4 (1991): 323-332.

[8] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In *Proceedings of KDD 1996*, 226-231. AAAI Press.

[9] González, M.C., Hidalgo, C.A., and Barabási, AL. "Understanding Individual Human Mobility Patterns." *Nature* 458, no. 7235 (March 12, 2009): 238-238.

[10] Greenwald, A., Hampel, G., Phadke, C., and Poosala, V., "An Economically Viable Solution to Geofencing for Mass-market Applications." *Bell Labs Technical Journal* 16, no. 2 (2011): 21-38.

[11] Huff, D. L. "A Probabilistic Analysis of Shopping Center Trade Areas*." Land Economics* (1963): 81-90.

[12] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. "Identifying Important Places in People's Lives from Cellular Network Data." *Pervasive Computing (2011)*: 133-151.

[13] Joseph, K., Tan, C. H, Afek, Y., Li, M. and Niu, K., "Beyond 'Local','Categories' and 'Friends': Clustering Users Based on Latent 'topics'", in *Proceedings of Ubicomp 2012,* 919-926

[14] Khoshgozaran, A., Shahabi, C. and Shirani-Mehr, H. "Location Privacy: Going Beyond K-anonymity, Cloaking and Anonymizers." *Knowledge and Information Systems* 26, no. 3 (2011): 435-465.

[15] Kures, M. Pinkovitz, Ryan, B., "Downtown and business district market analysis", http://fyi.uwex.edu/downtown-market-analysis/files/2011/02/Trade_Area_Analysis033011.pdf

[16] Li, N., and Chen, G. "Sharing Location in Online Social Networks." *Network*, IEEE 24, no. 5 (2010): 20-25.

[17] Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J. "I'm the Mayor of My House: Examining Why People Use Foursquare - a Social-driven Location Sharing Application." In *Proceedings of CHI2011*, 2409-2418. ACM, 2011.

[18] Liu, Z., Bonazzi, R., Fritscher, B., and Pigneur, Y. "Privacy-Friendly Business Models for Location-Based Mobile Services." *Journal of Theoretical and Applied Electronic Commerce Research* 6, no. 2 (August 2011): 90-107.

[19] MapInfo, Psyte U.S. Advantage. http://resource.mapinfo.com/static/files/document/1074655820383/psyte_brochure_us.pdf

[20] Martin, D., Alzua, A., and Lamsfus, C. "A Contextual Geofencing Mobile Tourism Service." *Information and Communication Technologies in Tourism* 201, 191-202.

[21] Munson, J. P., and Gupta, V. K. "Location-based Notification as a General-purpose Service." In *Proceedings of the 2nd International Workshop on Mobile Commerce*, 40-44, 2002.

[22] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. "An Empirical Study of Geographic User Activity Patterns in Foursquare." In *Proceedings of ICWSM'11* (2011).

[23] Partridge, K., and Begole, B. "Activity-Based Advertising." In *Pervasive Advertising*, edited by Jörg Müller, Florian Alt, and Daniel Michelis, 83-101. Human-Computer Interaction Series. Springer London, 2011.

[24] Provost, F., Martens, D. and Murray, A. "Finding similar users with a privacy-friendly geo-social design", http://www.everyscreenmedia.com/esmtwo/wp-content/uploads/2012/11/Finding_Similar_Users.pdf

[25] Qu, Y. and Zhang, J. “Regularly Visited Patches in Human Mobility”, to appear in *Proceedings of CHI 2013*, ACM

[26] Quercia, D., Lorenzo, G.D., Calabrese, F., and Ratti, C. "Mobile Phones and Outdoor Advertising: Measurable Advertising*." Pervasive Computing*, no. 99 (2011): 1-1.

[27] Reilly, W. J., *Methods for the Study of Retail Relationships*. University of Texas, Bureau of Business Research, 1929.

[28] Reilly, W.J. *The Laws of Retail Gravitation*, New York, Knickerbocker Press, 1931

[29] Roger, P. "Description of Consumer Spatial Behaviors: a New Approach*." International Journal of Research in Marketing* 1, no. 3 (1984): 171-181.

[30] Sliwinski, A. "Spatial Point Pattern Analysis for Targeting Prospective New Customers: Bringing GIS Functionality into Direct Marketing." *Journal of Geographic Information and Decision Analysis* 6, no. 1 (2002): 31-48.

[31] Song, C., Qu, Z., Blumm, N., and Barabási, AL. "Limits of Predictability in Human Mobility." *Science* 327, no. 5968 (2010): 1018-1021.

[32] Sun, A., Valention, J. and Seward, Z., A week on Foursquare http://graphicsweb.wsj.com/documents/FOURSQUAREWEEK1104/

[33] Tang, K.P., Lin, J., Hong, J., Siewiorek, D.P., and Sadeh, N. "Rethinking Location Sharing: Exploring the Implications of Social-driven Vs. Purpose-driven Location Sharing." In *Proceedings of Ubicomp 2010*, 85-94, ACM, 2010

[34] Ye, M., Yin, P., and Lee, W.C. "Location Recommendation for Location-based Social Networks." *In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 458-461, 2010.

[35] Zhang, D., Guo, B., Li, B., and Yu, Z. "Extracting Social and Community Intelligence from Digital Footprints: An Emerging Research Area." *Ubiquitous Intelligence and Computing* (2010): 4-18.

[36] Zhang, J., “Parcel-based high precision reverse geocoding", technical report, Pitney Bowes.

[37] Zhang, J., Qu, Y., and Teng, C.Y., “Understanding User Spatial Behaviors”, in *Proceedings of SRS 2013*, Rio, Brazil

[38] Zheng, Y., Zhang, L., Ma, Z., Xie, X., and Ma, W. Y. "Recommending Friends and Locations Based on Individual Location History." *ACM Transactions on the Web* (TWEB) 5, no. 1 (2011): 5.