Mining Travel Behaviors of Tourists with Mobile Phone Data: A Case Study in Hainan

Feng Ling[1,3], Tianyue Sun[1,3], Xinning Zhu[1,3], Qingqing Chen[2,3], Xiaosheng Tang[1,3], Xin Ke[4]
[1]Key Laboratory of Universal Wireless Communications Ministry of Education
[2]The Institute of Sensing Technology and Business BUPT, Wuxi, China
[3]Beijing University of Posts and Telecommunications
[4]Beijing Research Institute China Telecom BeiJing, China
Email: {lingfeng, sunty, zhuxn, cqq, txs}@bupt.edu.cn, kexin@ctbri.com.cn

***Abstract*—As the mobile devices are used widely, it is possible to research human mobility and behaviors by using the increasing mobile phone data. This paper aims to mine travel behaviors of users in Hainan, a popular tourist destination in China. The data is Call Detail Records (CDR) and Point of Interest (POI) information. Specifically, a data processing platform is developed. It can convert the large scale of CDR and location-based data to trajectories by utilizing the cross-domain data. Furthermore, a hierarchical analytical method is built on our platform to identify the difference of tourists respectively in temporal and spatial dimensions. Finally, the most popular travel patterns and top 3 travel routes for short-term tourists in Sanya are discovered from the trajectories of tourists according to the platform.**

Keywords-hierarchical analytical method; travel behavior; mobile phone data; platform

I. INTRODUCTION

In the past few years, the amount of mobile phones increases significantly, which is over 1.3 billion in 2016 in China. The widely used mobile phones can provide massive location data which facilitates the study of human mobility and behaviors. In fact, Call Detail Records (CDR) data and location-based data generated by telephone exchanges or other telecommunications equipments contain the geographic locations of every single communication activity, such as voice calls, short messaging service (SMS) and passive network events when the cell phone moves between network zones, or after some time of inactivity.

Human mobility is becoming a popular research topic because it relates to various fields such as traffic analysis, land use, social science and psychology. Some researches on tourists are based on travel survey data [4], however, such methods involve high costs and are not easily to scale up. In this paper, we make efforts to make use of the large scale CDR data and Point of Interest (POI) information to mine tourists travel behaviors, and Apache Spark is used for large-scale data processing. Generally, only tourists from other provinces are interested in insights of Hainan Province. So in this paper, we focus on the data of tourists from other provinces.

Different methods have been applied to mine the travel behaviors. However, because of the diversity of the tourists' behaviors in temporal and spatial dimensions, it is still difficult to capture and describe the travel behaviors of all tourists in Hainan. Tourists have different options in the number of days spent in Hainan and the number of cities they visited which makes it impossible to analyze travel behaviors in a single way. Besides, the trajectories that are extracted from records cannot reflect the real purposes of the tourists, for example, a tourist staying in a scenic area may have a meeting in a hotel rather than travel in the scenic area. It is necessary to add annotation for trajectories, which is helpful for understanding the travel behaviors.

In order to solve the problem above, a novel data processing platform is build in the paper. On the platform, the data processing part can deal with a larger scale of CDR and location-based data and the data analysis part can distinguish tourists by using a hierarchical analytical method from temporal and spatial dimensions. Specifically, we make efforts to mine tourists travel behavior from 521,588 tourists in Hainan and to find travel route patterns from their trajectories.

The paper is organized as follows. In Section II, we take a review of the related work. And in Section III, we introduce the proposed method which is used on our data processing platform. A case study is presented in Section IV. And Section V gives the conclusion and future research directions.

II. RELATED WORKS

In the field of human mobility, many efforts have been made to discover the pattern of the crowd and various types of data have been used in the analysis, such as CDR, GPS and WiFi. By using CDRs, researchers are able to analyze human mobility and patterns. Shan Jiang et al.2013[1] reviewed the urban computing for mobile phone traces. The article focused on identifying users' stay points and discovering users' frequent daily motifs, which inspired our research. To identify the stay points of users, Peiyu Yang et al. 2014 [2] extracted users' trajectory from CDRs and then presented a method to select significant places, such as home and work places. And there is also a hierarchical analytical model presented by Fahim Hasan Khan et al. 2015 [3] in which the CDR data is used to find facts on the daily life activities of urban users in multiple layers. And in the field of human daily motifs, Christian M. Schneider et al. 2013 [4] analyzed the human mobility and calculated the most frequent motifs by using travel survey data. The result showed the motif distributions in different cities were

978-1-4673-9026-2/16/$31.00 ©2016 IEEE

similar. Christian M. Schneider et al. 2013 [5] introduced the rules that described each motif in details. Combining with the city planning, Chaogui Kanga et al. 2012 [6] mined human mobility patterns on the level of urban morphology perspective, and came to the conclusion that the human mobility is influenced by the urban areas and urban layout.

Many other types of data are also used to researches. Andrea Cuttone et al. 2014 [7] studied the feasibility of inferring human mobility from sparse, low accuracy mobile sensing data. Their results suggested that low-resolution data allows accurate inference of human mobility patterns. Rui Wang et al. 2015 [8] used passive sensing data and self-reports from students' smart phones to understand individual behavioral differences between students with good and bad academic performances during a single 10-week term. Geert Vanderhulst et al. 2015 [9] proposed the use of WiFi probes, management frames of WiFi, that periodically radiate from mobile devices, and existing WiFi access points to automatically capture radio signals and detected human co-presence.

In the field of travel behavior, Mingqiang Xue et al.2014[10] identified the tourists among public commuters using the public transportation data provided by Singapore's Land Transport Authority and then revealed the travelling patterns of tourists. Santi Phithakkitnukoon et al. 2014 [11] described a framework that capitalizes on the large-scale opportunistic mobile sensing approach for tourist behavior analysis.

III. METHODOLOGY

In this section, we introduce the method which is used on our platform. In the data processing part, after data cleaning, we identify stay points and add annotation to stay points which converts the large scale records to trajectories with semantics. Then, in the data analysis part, we propose a hierarchical analytical method to analyze travel behaviors from temporal and spatial dimensions by using the features extracted from CDR and location-based data.

A. Data Processing

1) Data set

In our analysis, CDR and location-based data provided by one of the largest telecommunications operators in China, contains more than 10 millions anonymized mobile phone records from users in Hainan province for a period of two months in 2015. To deal with the large-scale data, we use Apache Spark, a cluster computing framework for big data processing.

The data set consists of CDR records and location-based data. The CDR records contain the details of a telephone call or other communications transaction which includes the base station's identifier, the calling and called number, and the timestamps corresponding to the transaction. And the location data can be produced when a user hand over from one cell to another or after 30 minutes of inactivity. In addition, we also have the geographical position information (latitude, longitude) of most base stations which could be regarded as the location of the users. From the CDR and location-based data, we can easily extract the users' stay points and trajectory. In this paper, the data processing flow is shown as Fig. 1.

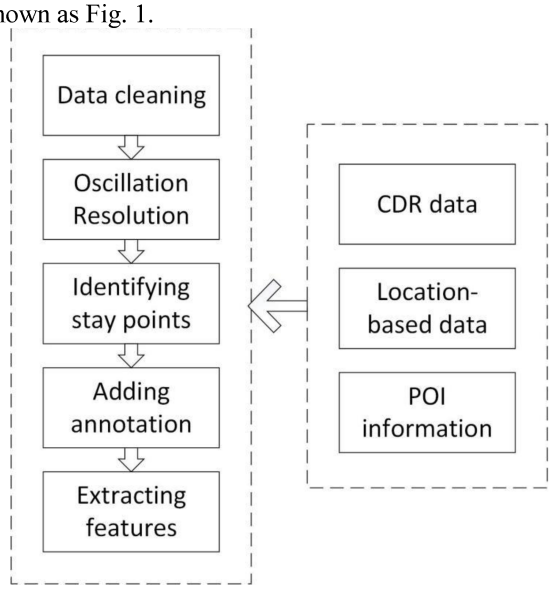

Figure 1. Data processing flow chart in the data processing part

2) Data cleaning and oscillation resolution

This paper focuses on users who are not local residents, the users whose calling number does not belong to Hainan province are extracted as tourists. In addition, a few users who have too many records are filtered out, because these are likely to be some service numbers and may lead to misunderstandings of the users' travel behaviors. Besides, we discard the users who have few records, which may not contribute to the analysis.

When a mobile phone is in the overlapped areas between adjacent cells, it may switch between two cells but actually the user's location hasn't changed. To eliminate such noise, we refer to [12] for detecting and removing oscillation records.

3) Identify stay points

A stay-point is identified by a sequence of consecutive cell phone records limited by both temporal and spatial constraints. The rules are as follows: when two single points are within a certain distance D and a certain time interval T, the two points will be joined together as one stay point. The value of D and T is set to be 200 meters and 10 minutes respectively. After finishing this step, a user's trajectory is constructed by a series of stay points $(p_1, p_2, \ldots, p_n)$. $p_i = (x_i, y_i, t_i)$, x_i and y_i are the longitude and latitude of p_i. t_i is the time when the user stayed in p_i.

B. Adding Annotation

In this study, we use a network crawler to collect POI information in Hainan Province from the Internet. Taking Sanya city for example, the amount of POIs in Sanya is 17,758, and the categories of POIs contains caters, schools, hotels, residences, tourist attractions and so on. And each POI has its position information (latitude, longitude).

According to these information, we can identity what kind of the stay point is in the records. A stay point may belong to "hotel", "attraction", "home", or "unknown", and the trajectory is consist of these meaningful stay points. How to add a semantic annotation will be discussed in the next section.

C. Hierarchical Analysis

We developed a hierarchical method on our platform to mine tourists travel behavior from temporal and spatial perspective as Fig. 2. The method is consist of five layers. We input all users into the first layer with their features. Based on the temporal analysis, all users are classified into short-term users and long-term users. The second layer is dependent on the previous layer. It classifies short-term users into two groups by the number of cities they visited. In the third layer, the method only picks users staying in Sanya and makes spatial analysis. As the Fig. 2 shows, after all five layers' filtering, we can gain the classification results of users. However, in order to make temporal analysis, firstly we should identify significant stay points from users' trajectories.

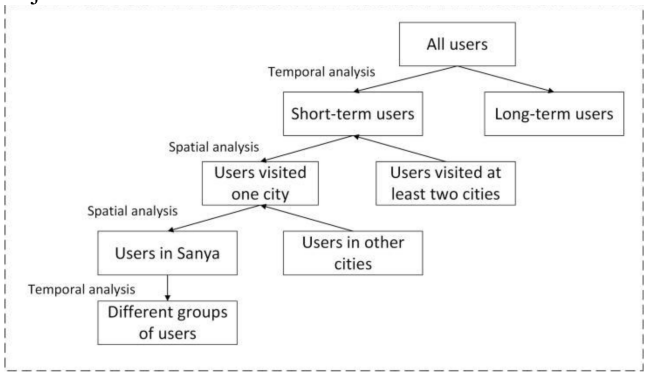

Figure 2. The hierarchical method in the data analysis part

IV. Case Study in Hainan

This section presents a case study about the travel behavior of Hainan tourists, using our data processing platform. It starts with feature extraction, followed by the analysis and results. At last, we study the travel preference and discover trajectory patterns and the top 3 popular travel routes in Sanya.

A. Feature Extraction

Hainan is a well-known tourist destination, located at the southernmost part of China. A large number of people travel there for tourism, business, or visiting their relatives and friends every year. The behaviors of tourists are varied by the purpose of their travel. So we define some features which are helpful to our analysis of tourists' behaviors.

1) The number of days a tourist spent in Hainan: The behaviors of the short-term tourists and the long-term tourists are quite different. So the first step is to count how many days each tourist stayed in Hainan. The results of August and December are shown in Fig. 3. It can be seen that the number of tourists in December, which is the peak season in Hainan, is much lager than that in August. Most of the short-term tours last no more than seven days, and a five-day tour is the most popular one. According to the tourism websites, many five-day tours are provided to tourists, which can partly explain the result. There is also a peak value of 31 days which corresponds to the long-term tourists. They might stay there more than a month, but because we use one-month records, the tourists staying for 31 days are regarded as long-term tourists. Actually, a lot of northern people in China choose Hainan for a long vacation to escape the bitter cold in winter.

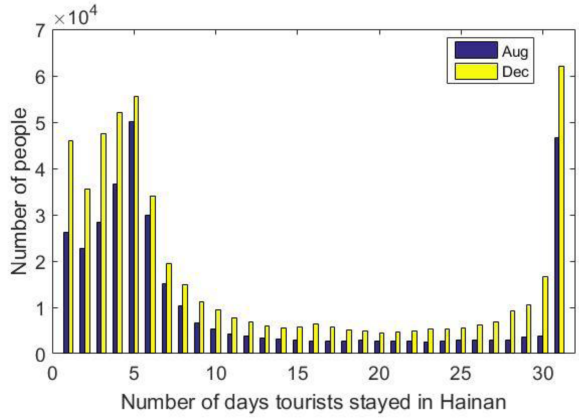

Figure 3. Comparison of the number of tourists in August and December

2) The maximum distance between two locations a tourist visited: It represents the maximum span of the area covered.

$$D = \max_{i,j \in \{1,2,\ldots n\}} d(P_i, P_j) \tag{1}$$

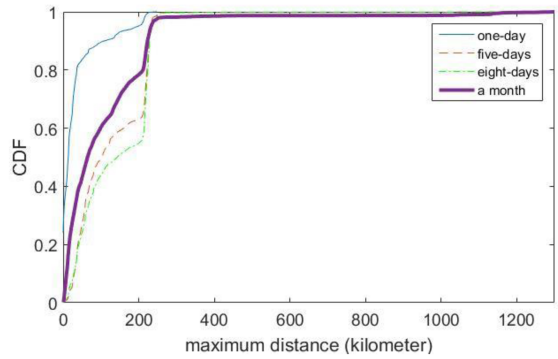

Figure 4. CDF of the maximum distance

The CDF curves with the maximal distance in different travel time in Hainan are shown in Fig. 4. It shows that the shapes are different. For the short-term tourists, it is obvious that a turning point appears at about 200 kilometers, and the slope becomes large. It suggests that a part of tourists traveled across Hainan province in their tours. After looking up them, we discover that they traveled from Haikou, which is located at the northern coast of Hainan, to the well-known tourist destination on the southern coast, Sanya. The percentage of tourists who traveled across Hainan is growing with the number of days they spent in Hainan except the long-term tourists. Because long-term tourists are likely to

settle in one city to take a vacation, they just traveled within the range of the city.

3) The radius of gyration of a tourist: It reflects a tourist's activity area and is defined as the deviation from the centroid of the places the user visited. The contribution of each place is weighted by the time the user spent in that place. We denote the time as t_i. Let p_{center} represent the centroid of the places a tourist visited so the radius of gyration of a tourist can be expressed as

$$R = \sqrt{\frac{1}{\sum_{i=1}^{n} t_i} \sum_{i=1}^{n} t_i \bullet d\left(p_i, p_{center}\right)} \quad (2)$$

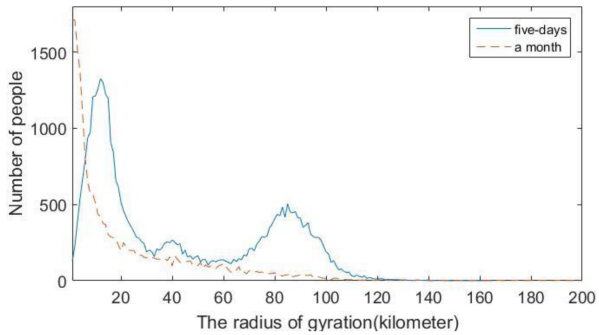

Figure 5. The radius of gyration of different types of tourists

Fig. 5 shows the comparison between five-day tourists and long-term tourists. The blue curve which represents five-day tourists has two peaks. The second appears at about 85 kilometers, which is caused by tourists who traveled from one city to another. The red one representing long-term tourists which shows a decreasing trend, becomes more close to X axis when X is over 20 kilometers. We can conclude that the short-term tourists prefer to traveling, while the long-term tourists, on the contrary, behave more like local residents. The radius of their activities is small because such tourists settle in the city and usually have a fixed house to stay, their daily lives are the same as local residents.

4) The number of cites a tourist visited: Tourists traveling to several cities behave differently from others that stay in one city. To simplify the analysis, we focus on tourists staying in Hainan for five days. If a tourist spend more than 10 percent of the travel time in one city, we consider him or her actually visiting the city instead of passing. The result shows that in all five-day tourists, 54 percent of them stayed in only one city, and 38 percent of them visited two cities, only 9 percent of them visited more than two cities.

5) The percentage of time spent in scenic areas of a tourist in the daytime: The daytime is defined as the period from 8am to 5pm and we calculate the time percentage of stay points which are labeled as “attraction” in the daytime.

6) The percentage of time spent in hotels of a tourist in the daytime: We calculate the time percentage of stay points which are labeled as “hotel” and “home” in the daytime.

B. *Identify Tourist Attractions and Hotels*

To study tourists' behavior, we calculate the time percentage they stay in tourists attractions and in hotels during the daytime. Using the POI information in Sanya, we respectively make a list for scenic areas and hotels. The list of scenic areas contains all 4A and 5A scenic spots as well as most of parks and beaches. The list of hotels contains hotels and residential areas in Sanya. For each tourist, we extract the stay points in the daytime and calculate the distance between the stay points and the POIs. When the distance between them is under 0.5 kilometers, the stay point will be marked as the type of that POI. By doing this, we can classify stay points into “hotel” or “attraction” or “unknown”.But we find it is hard to categorize a stay point when it belongs to a hotel which is located in the scenic area because it has two attributions of POI, i.e. hotel and attractions. To solve this problem, we need to identify users' home.

C. *Identify Tourists Home*

For each tourist, we pick the stay point where the tourist spent the longest time at night as home. Since we pick the information on daily basis, we ignore the problem that a user may live in different places at different night. Moreover, when a stay point belongs to two different types of POI, we will classify it into “hotel” if the user stay here that night. Otherwise, it will be classified into attraction. In the next step, we calculate the time percentage that users spent in visiting scenic area or staying in a hotel during the whole five days. Fig. 6 shows the result.

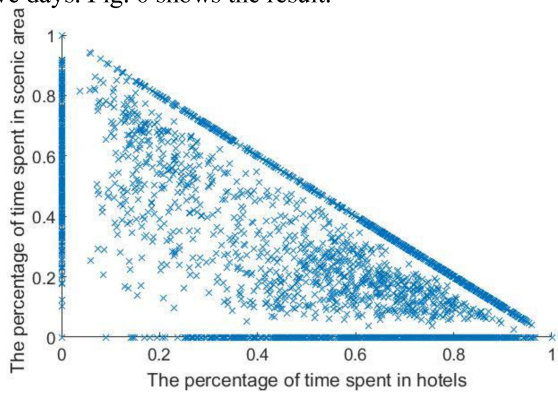

Figure 6. Scatter plot of tour time and hotel time, the X axis represent the percentage of time in hotel, the Y axis represent the percentage of time in attractions.

It can be seen that a number of tourists stay in the hotel or at home during the entire daytime. We think these tourists may be on a business trip or visit their friends, so they stayed in hotels to have a meeting or just stayed at home and didn't spend much time in traveling. There is another kind of tourists who spent the entire daytime in the scenic areas. These tourists must come to Hainan for a tour. Morever, many tourists stayed not only in scenic areas but also in hotels in their daytime. These users are divided into two groups in Fig. 7.

In Fig. 7, six users are divided into two groups, and the cumulative distance of them is quite different. The users

whose cumulative distance is high are likely to be tourists because they spent more than 60 percent of their daytime in the scenic areas. Their cumulative distance increased fast along with the cumulative time, which indicates that they visited several scenic areas in different places. Another three tourists represented by dotted lines have a short cumulative distance in Sanya. They stayed in a place for a long time but moved little. These three users spent only 15 percent of their daytime in the scenic areas but 80 percent at home. We think these tourists may be on a business trip or visit their friends, so they have little time to visit scenic areas.

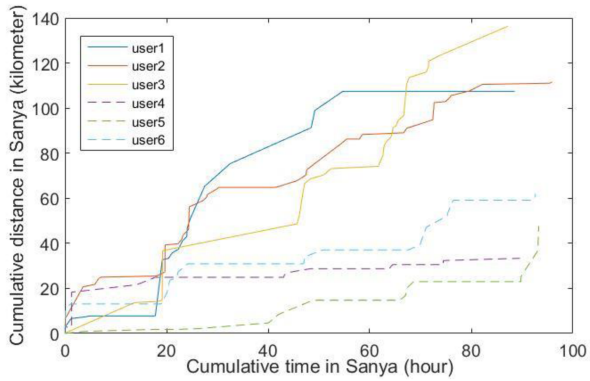

Figure 7. Comparison of two users' cumulative distance from different groups

D. *Travel Preference*

According to the trajectories, we do the research of patterns mining. Each tourist's trajectories in five days may be different. However, there exists common patterns shared by a part of tourists that represent popular scenic areas or hot travel routes. We define a trajectory as a series of scenic areas a tourist visited in the daytime. Tourists who visited the same places must have the common patterns in trajectories. In this study, we discovered the most common trajectory patterns of tourists in Sanya. The pattern may be a single place that most tourists visited in their five-day tour, or a route that consists of several scenic spots for short-term tourists. Therefore, the most common trajectory patterns wil not be very long.

Fig. 8 shows the top 3 popular scenic areas in Sanya. The most popular scenic area is Yalong Bay, followed by the second one, Sanya Dadonghai, and the third one is Sanya Nanshan. All of the spots are beside the sea. Fig 9 shows the top 3 travel routes among two scenic areas. The most popular route is between Tianya Haijiao and Sanya West Island. The top 2 route is from Yalong Bay, which is in two different trajectory patterns, to Wuzhizhou Island and the top 3 route is from Sanya Nanshan to Yalong Bay.

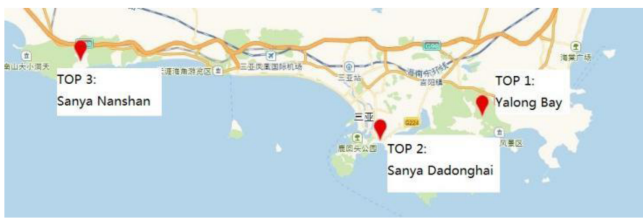

Figure 8. Top 3 scenic areas

Figure 9. Top 3 travel routes

V. CONCULSION

In this work, we make an effort to mine travel behaviors by using CDR data and POI information. Compared with the travel survey data collected by questionnaires, our data has a much larger scale and costs less. Using the data processing platform, we convert the large scale of CDR and location-based data to trajectories. Then we present a case study in Hainan using our data processing platform. The result presents the different groups of tourists in Sanya, which proves the possibility of mining travel behaviors through mobile data. Besides, we have discovered trajectory patterns and the top 3 popular travel routes in Sanya. In our future work, we plan to analyze the similarity of tourists' trajectory, and try to discover tourists' mobility motifs by the data. We will also make efforts to improve the performance of our platform.

ACKNOWLEDGMENT

This work was partially supported by the National High Technology Research and Development Program of China (Grant No. 2014AA01A706), National 3rd Key Program project (No. 2014ZX03002002-004), Funds for Creative Research Groups of NSFC (Grant No. 61421061) and the Fundamental Research Funds for the Central Universities (2014ZD03-01).

REFERENCES

[1] Jiang Shan, Fiore Gaston A., Yang Yingxiang, Ferreira Jr. Joseph, Frazzoli Emilio, González Marta C.:A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2013)

[2] Yang Peiyu, Zhu Tongyu, Wan Xuejin, Wang Xuejiao: Identifying Significant Places Using Multi-day Call Detail Records, ICTAI (2014)

[3] Khan Fahim Hasan, Ali Mohammed Eunus, Dev Himel: A Hierarchical Approach for Identifying User Activity Patterns from Mobile Phone Call Detail Records, NSysS (2015)

[4] Schneider Christian M., Rudloff Christian, Bauer Dietmar, González Marta C: Daily travel behavior: Lessons from a week-long survey for the extraction of human mobility motifs related information,UrbComp (2013)

[5] Schneider Christian M., Belik Vitaly, Couronné Thomas, Smoreda Zbigniew, González Marta C: Unravelling daily human mobility motifs, Journal of the Royal Society Interface, v 10, n 84, July 6, 2013

[6] Kang Chaogui, Ma Xiujun, Tong Daoqin, Liu Yu: Intra-urban human mobility patterns: An urban morphology perspective, Physica A: Statistical Mechanics and its Applications, v 391, n 4, p 1702-1717, February 15, 2012

[7] Cuttone Andrea, Larsen Jakob Eg, Lehmann Sune: Inferring Human Mobility from Sparse Low Accuracy Mobile Sensing Data, UbiComp (2014)

[8] Wang Rui, Harari Gabriella, Hao Peilin, Zhou Xia, Campbell Andrew T: SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students, UbiComp (2015)
[9] Vanderhulst Geert, Mashhadi Afra, Dashti Marzieh, Kawsar Fahim: Detecting Human Encounters from WiFi Radio Signals, MUM (2015)
[10] Xue Mingqiang, Wu Huayu, Chen Wei, Ng Wee Siong, Goh Gin Howe: Identifying Tourists from Public Transport Commuters, KDD (2014)
[11] Phithakkitnukoon Santi, Horanont Teerayut, Witayangkurn Apichon, Siri Raktida, Sekimoto Yoshihide, Shibasaki Ryosuke: Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan, Pervasive and Mobile Computing, v 18, p 18-39, April 1, 2015
[12] Wu Wei, Wang Yue, Gomes Joao Bartolo, Anh Dang The, Antonatos Spiros, Xue Mingqiang, Yang Peng, Yap Ghim Eng, Li Xiaoli, Krishnaswamy Shonali, Decraene James, Nash Amy Shi: Oscillation Resolution for Mobile Phone Cellular Tower Data to Enable Mobility Modelling, Proceedings - IEEE International Conference on Mobile Data Management, v 1, p 317-324, October 5, 2014