Checking In or Checked In: Comparing Large-Scale Manual and Automatic Location Disclosure Patterns

Eric Malmi
Idiap Research Institute
Martigny, Switzerland
eric.malmi@idiap.ch

Trinh Minh Tri Do
Idiap Research Institute
Martigny, Switzerland
do@idiap.ch

Daniel Gatica-Perez
Idiap and EPFL
Switzerland
gatica@idiap.ch

ABSTRACT

Studies on human mobility are built on two fundamentally different data sources: manual check-in data that originates from location-based social networks and automatic check-in data that can be automatically collected through various smartphone sensors. In this paper, we analyze the differences and similarities of manual check-ins from Foursquare and automatic check-ins from Nokia's Mobile Data Challenge. Several new findings follow from our analysis: (1) While automatic checking-in overall results in more visits than manual checking-in, the check-in levels are comparable when visiting new places. (2) Daily and weekly check-in activity patterns are similar for both systems except for Saturdays – when manual check-ins are relatively more probable. (3) A recently proposed rank distribution to describe human mobility, so far validated on manual check-in data, also holds for automatic check-in data given a slight modification to the definition of rank. (4) The patterns described by automatic check-ins are in general more predictable. We also address the question of whether it is possible to find matching places across the two check-in systems. Our analysis shows that while this is challenging in areas such as city centers, our method achieves an accuracy of 51 % for places that are not homes of phone users.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous; J.4 [**J.4 Social and Behavioral Sciences**]: Sociology

General Terms

Experimentation, Human Factors

Keywords

check-ins, Foursquare, MDC, place matching

MUM '12, December 04 - 06 2012, Ulm, Germany

1. INTRODUCTION

Location, as a key contextual cue in ubiquitous computing, has been a target in research and industry for years. With the advent of smartphones and big mobile data, one could argue that daily life places are (literally and metaphorically speaking) finally on the map, either automatically discovered from mobile sensors [12, 15, 11, 24] or manually disclosed and shared in location-based social networks (LBSNs) [22, 18, 5]. As a result, a surge of work in computational social science [17] is using large-scale mobile data, where instantaneous location is disclosed manually or automatically, to address questions related to the existence of general patterns in human mobility and of limits on location predictability [10, 20]. For these analyses, most existing works rely on *individual sources of location traces* – automatic check-ins estimated from cell tower data or smartphone sensors, or manual check-ins from LBSNs like Foursquare (4sq). Furthermore, these works assume that the input data accurately represents human mobility, somewhat independently of the method used to generate check-ins.

This assumption, in our opinion, needs to be examined, as there are fundamental differences in how check-ins are produced. Manual check-in systems (like Foursquare, where people freely choose which places they want to check into) are driven by individual preferences, specific incentives, and bounded by people's interest and attention. A few studies [18, 5] have investigated what motivates LBSN users to check in to certain places, and how people's self-representation choices affect location sharing decisions. For instance, it was found in [18] that some users do not check in at places they consider boring, such as home, while others do not share their visits to fast food restaurants as they consider those embarrassing. On the other hand, people who value more the gaming aspect of 4sq also check in at their homes in order to become the mayors of their homes. In contrast to the subjective nature of LBSN check-in patterns, automatic check-in systems (in which people's location is inferred from GPS trajectories and WLAN access points) are agnostic to the above issues and provide objective information about users' whereabouts, but often suffer from their own limitations including sensor failures and battery constraints.

In this work, we examine this question in detail, by systematically comparing large-scale data produced by manual and automatic location disclosure systems. We are not aware of previous studies addressing this goal on longitudinal, country-level data. As an instance of a manual check-in

system, we use data from Foursquare [22]. As an example of an automatic check-in system, we use data from the *Nokia Mobile Data Challenge* (MDC) [16], which in turn originates from the Lausanne Data Collection Campaign (LDCC) [14], where phones used a variety of sensor data to infer instantaneous locations and visited places. The study is conducted using data collected over several months in Switzerland.

Our paper has two main goals. The first one is to study how the overall behavior of 4sq users and MDC users reflects the different nature of manual and automatic check-in systems. Our analysis reveals significant differences in amounts of check-ins and place predictability across systems, but also finds many similarities and some nuanced differences w.r.t. periodic check-in patterns and rank distributions. The findings of this comparison contribute to the understanding of how well previously reported patterns [10, 20] can generalize across check-in systems.

A challenge in automatic check-in systems is to identify the exact place a user is at given a set of nearby places [18]. Our second goal is to address this question by quantifying to what extent corresponding places between the two data sets can be matched. Another motivation for matching places across data sets is that it would allow to use data sets in complementary manners, e.g., to learn specific features of manual checked-in places to enrich automatic checked-in data and vice versa.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the two data sets used in our study. Section 4 presents our analysis on visiting behavior from the perspectives of manual and automatic check-ins. Section 5 presents and discusses our work on matching places across check-in data sets. Finally, conclusions and future directions are discussed in Section 6.

2. RELATED WORK

Automatic check-in data can be obtained from various sources. The reality mining data set [6] which is more limited but bears similarities with the MDC data set has been widely analyzed among researchers (e.g. [2, 7]). Coarse resolution automatic check-in data can be obtained from cell tower information. The mobility patterns found in this type of data have been studied, e.g. by Gonzalez et al. [10]. Another way to collect information about human mobility is to follow the circulation of bank notes which has been done in [1]. Instead of raw GPS trajectories, it is more convenient to study visit sequences between places which can be extracted from GPS traces and/or other data types such as Wi-Fi/GSM radio. In the place learning literature, there are several definitions of place, such as a small circular region [12], Wi-Fi/GSM fingerprint [15, 11, 13], or a multivariate Normal distribution [21]. Recently, Zheng et al. [24] proposed a two-stage algorithm that first detects stay points from GPS traces and then clusters these into places. The algorithm was later improved by Montoliu and Gatica-Perez [19] by taking into account other sensors such as Wi-Fi. In the MDC data, the visit sequences have been extracted using the improved version [16].

Several mobile applications, including Foursquare, Whrrl and Gowalla, collect manual check-in data. Lindqvist et al. [18] conducted interviews and two surveys in order to understand what motivates people to use Foursquare to share their locations. They reported several reasons why people use it including a gaming aspect, keeping in touch with friends, and discovering new places. Cramer et al. [5] also conducted interviews and a survey in order to gain insight into Foursquare usage and they were able to confirm the findings by Lindqvist et al. Additionally, they looked into the demographics of people's Foursquare friends. Cheng et al. [3] collected a data set of 22 million manual check-ins mainly from Foursquare but from several other applications as well. One thing they looked into was the distribution of displacement distances. Later on, Noulas et al. [20] showed that rather than the distance of a displacement, it is the rank of the displacement what characterizes its probability more accurately. Another aspect Cheng et al. studied was the daily and weekly check-in patterns of the users. Ye et al. [23] looked into this type of patterns in Whrrl data, but in addition they analyzed the daily and weekly check-in patterns of different types of places, showing that places can be characterized not only by their user-assigned category tags, but also by their temporal check-in patterns.

Much work has been done to characterize automatic and manual location disclosure systems separately. To our knowledge, the only work that has used both automatic and manual check-in data is by Cho et al. [4]. The objective of their study is to look at specific human mobility patterns and see if those can be found across datasets. In contrast, the goal of our study is to look at the differences and similarities of manual and automatic check-in systems on a general level and to study in what sense the two systems complement each other.

3. DATA SETS

In this section, we present the MDC data set, which contains automatic check-in data, and the 4sq data set, which contains manual check-in data. For notational convenience, *check-in* and *visit* are used interchangeably hereinafter. Both data sets were collected in Switzerland.

3.1 Automatic Check-In Data

The automatic check-in data comes from Nokia's Mobile Data Challenge[1], described in [16], which originates from the Lausanne Data Collection Campaign (LDCC) [14]. The MDC Dedicated Track data set contains daily life data from 80 users and about 16 months (Sep 2009 – Feb 2011). The population is concentrated in the Suisse Romande region, the French speaking part of Switzerland, but their data was recorded wherever they were inside the country. Users are a combination of students and professionals, mainly in the 22–33 age range.

The volunteers in LDCC were given cell phones that automatically recorded various types of sensory information, including GPS, WLAN, Bluetooth, cell tower, and accelerometer data, and other types of information, such as call logs and application usage logs. The MDC data set also contains *place visit sequences* that were automatically inferred based on the GPS trajectories and WLAN access points. In this work, we only use the place visit sequences. They contain

[1] `http://research.nokia.com/page/12000`

the identifier labels of the visited places and the start and end times of the visits. A visit is defined to have a minimum duration of 20 minutes meaning that if a user stays in a place for less than 20 minutes, he/she is considered to be on the move. This implies that many actual places in daily life, involving short stays like bus stops, metro stations, etc. are likely not included in the data.

The algorithm for detecting the visited places is described in [19] and consists of two stages: In the first stage, the temporally consecutive location points of a user are grouped into *stay points*. Then in the second stage, the stay points are grouped into *stay regions*, i.e. the places. Each place has a pair of coordinates, which corresponds to the center of mass of the stay points, and a radius of 100 meters, which is the maximum distance from the center to the location points. We run the place detection algorithm for each user separately, so that each user has associated a unique set of places. Note that the resulting location for each place discovered by this algorithm (latitude and longitude) was not part of the original MDC Dedicated Track data.

Some places have been given semantic labels by the users. The users have been asked to select labels for their most frequently visited places and some of the rarely visited places from a fixed list of category labels, such as home, work, restaurant, etc.

3.2 Manual Check-In Data

The manual check-in data comes from *Foursquare*[2] (4sq) which is a highly popular location sharing application. The data is collected through Twitter as some 4sq users have allowed their check-ins to be published on their Twitter stream, which allows the collection of longitudinal data per user. Similar techniques to collect 4sq data have been used in [3, 20]. For comparison purposes, we focus on the country-level check-ins from Switzerland, where the MDC data has also been collected.

We have collected the 4sq data between December 19, 2011 and June 21, 2012, containing 12882 different users in total. However, most of these users have only a couple of check-ins possibly corresponding to an initial interest and therefore we select *active users* who have made at least 40 check-ins and whose first and last visit are at least 4 weeks apart from each other. These choices result in 302 active users who are used for Section 4. The check-ins consist of a latitude, longitude, place title, and for most places a place category[3].

4sq check-ins and MDC visits have three fundamental differences: The first and the main difference is that 4sq check-ins are manual. When a 4sq user wants to check-in he/she gets a list of nearby places from which the user manually selects the suitable one (illustrated in Figure 1). Thus for most 4sq users the data is more sparse as the users do not often check in at every place they go to. In addition, our 4sq data set has been collected through Twitter, and only a fraction of 4sq users share all or part of their check-ins this way (also note that most studies using 4sq data have also obtained the data through the same procedure.) The second key difference relies on the definition of a place. While MDC places correspond to circular regions, 4sq places are physical coordinates suggested by the system which are more accurate in terms of precision. The third difference is that the MDC visits have both a start and an end time, whereas the 4sq visits have only the time of checking in, which is not necessarily in the beginning of the visit.

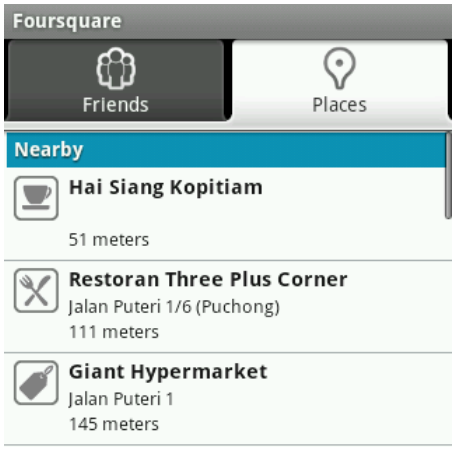

Figure 1: Screenshot from the Foursquare application where the user selects the place where he/she wants to check into from a list of nearby places. The photo was taken from `http://www.askabouthugo.com/wp-content/uploads/2010/05/foursquare-2.png`.

Some basic statistics regarding both data sets are shown in Table 1 and the locations of all places are shown in Figure 2. While the 4sq places (in red) distribute rather evenly over whole Switzerland, the MDC places (in blue) are centered in the French speaking part of Switzerland and especially around Lake Geneva where the MDC population lives. We have also included the places visited by inactive users. In total, there are 7281 MDC places and 17482 4sq places.

Table 1: Basic statistics of the MDC and 4sq data sets. *#labeled places* is the number of places for which the users have given a semantic label. *#active users* is the number of users with at least 40 visits and one month of data in total and *#visits* is the number of visits made by the active users.

	MDC	4sq
#places	7281	17482
#labeled places	398	16382
#users	80	12882
#active users	80	302
#visits	51607	40629
First visit	30 Sep 2009	19 Dec 2011
Last visit	4 Feb 2011	21 Jun 2012

4. VISITING BEHAVIOR

In this section, we compare the visiting behavior in MDC and Foursquare data. In Sec. 4.1 and Sec. 4.2, we examine two fundamental features of visits, namely: how often and

[2] `https://foursquare.com/`

[3] List of categories can be found in: `http://aboutfoursquare.com/foursquare-categories/`

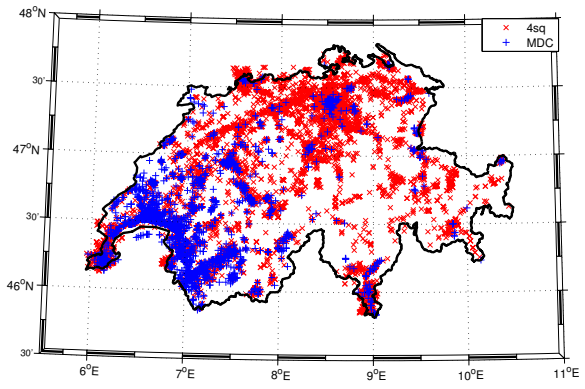

Figure 2: All MDC and 4sq places on a map of Switzerland. Note that in the Lake Geneva area, the two data sets are overlapped.

when people visit places. In Sec. 4.3 and Sec. 4.4 we study two additional aspects related to sequences of visited places.

4.1 How Often Do People Visit Places?

Figure 3 shows the distributions of the average number of visits per day (*vpd*) which are quite different for the two data sets. For most MDC users, *vpd* is between 2-4 (median = 3.1) and no user has a $vpd > 6$. On the other hand, more than half of the 4sq users have $vpd < 2$ (median = 1.7) but there is also one 4sq user who has as many as 19 *vpd*. Two-tailed t-test with unequal variances at 99 % confidence level confirms that *vpd*s of MDC and 4sq users have different means.

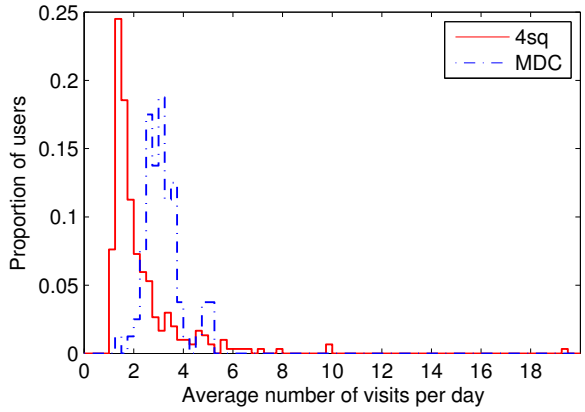

Figure 3: Empirical distributions of the average number of visits per day for manual and automatic systems.

This shows an obvious advantage of automatic check-ins: many 4sq users check in only occasionally, whereas MDC users record new visits regularly if they are carrying their mobile device with them. However, in the MDC data, a visit has been defined to last for at least 20 minutes. Thus, shorter visits will not be recorded, which might explain why the MDC data lacks a long tail, i.e., users with numerous visits per day which can be found in the 4sq data.

Let us also study if there are differences in the visiting behavior among 4sq users. We define three user categories based on their average check-in activity: users with less than 1.5 visits per day (*vpd*) are denoted by "low", users with $1.5 \leq vpd \leq 4.5$ by "mid", and users with $vpd \geq 4.5$ by "high". In Table 2, we have calculated statistics of the weekly visit frequencies between the two data sets and between different 4sq user categories. The table shows how many visits people record per week, and how many distinct places and new places they record per week on average[4].

Table 2: Weekly averages for number of visits, distinct places, and new places visited during the week. Last column shows the number of users in each category. 4sq-* refer to activity categories of 4sq users with different amounts of visits per day.

	#visits	#distinct	#new	#users
MDC	17.0 ± 5.6	6.7 ± 2.0	3.2 ± 1.2	80
4sq	7.0 ± 7.6	4.7 ± 4.2	3.1 ± 2.6	302
4sq-low	2.7 ± 0.9	2.3 ± 0.7	1.7 ± 0.6	97
4sq-mid	7.0 ± 4.5	4.8 ± 3.0	3.3 ± 2.0	182
4sq-high	25.4 ± 13.3	13.9 ± 7.1	8.3 ± 4.4	23

MDC users record over twice as many (143 % more) visits per week and 43 % more distinct places than 4sq users but, quite interestingly, 4sq users record roughly the same number of new places. How to interpret these results? On one hand, the larger number of visits for the automatic case is a clear reflection of the fact that this system has no burden on human memory or attention compared to the manual case. But there are other reasons related to checking in at home or work. The MDC users who have labeled their home or workplace make on average 24 % and 17 % of their visits to these places, respectively. In contrast, a survey conducted by Lindqvist et al. [18] showed that a vast majority of 4sq users never check in at their homes and less than half check in at their workplace on a daily basis. On the other hand, the fact that the overall number of new places are similar for both cases highlights the novelty-driven feature of location sharing systems, i.e., users are motivated to make the explicit effort of checking in, while for an automatic system, depending on its design, a new place might be treated just like previously seen places in terms of detection.

The numbers of distinct places and new places visited on average per week are shown for each user colored according his/her activity category in Figure 4. There is clearly a correlation between these two variables which means that people who go to many distinct places per week also visit many new places per week. Both Table 2 and Figure 4 show that there are very different types of 4sq users in terms of activity levels.

[4]The rate at which people record new places decreases slightly as a function of time since the number of potential new places to explore decreases. In order to make the *new places per week on average* comparable between users with several months and users with only a couple of months, we reset the set of visited places to the empty set every two months.

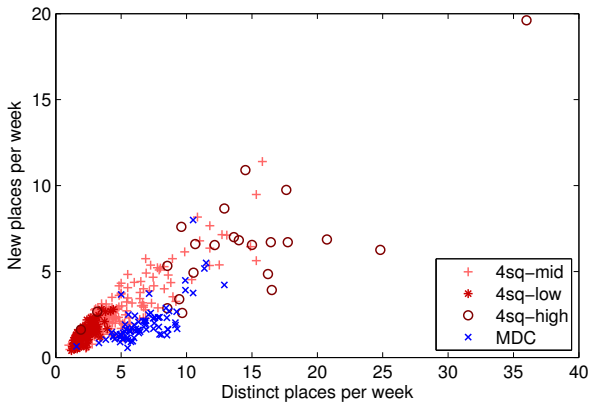

Figure 4: Each user's average number of distinct and new places visited per week. The correlations between the two variables are 0.82, 0.93, 0.88, and 0.87 respectively for 4sq-low, 4sq-mid, 4sq-high, and MDC users.

4.2 When Do People Visit Places?

Figure 5 shows the distributions of MDC and 4sq visits over hours of the day and days of the week. The hour distributions have three peaks corresponding to the arrival to work between 8-10^{h}, lunch break around 12-14^{h} and some after-work activity between 18-19^{h}. Furthermore, MDC visits peak in the night between 3-4^{h} which is due to a daily reset of the phone that causes artificial visits [14].

Cheng et al. [3] have found the same pattern of three check-in peaks for other cities, namely Los Angeles, New York City, and Amsterdam. For these cities the after-work peak is always the highest. Interestingly, in the 4sq and MDC data from Switzerland, the highest peak is around noon. This can be explained by cultural differences since in Switzerland the shops typically close earlier than many other countries (18:30^{h} during weekdays) and so urban activity decreases.

The distributions over days of the week on the right hand side of Figure 5 are also rather similar except for Saturday, when there are more 4sq visits than MDC visits. This interesting difference can be explained by looking at Figure 6, which shows the time distributions for the 4sq user categories (see Sec. 4.1) over hours of the day and days of the week. The highly active users check in relatively less frequently in the morning than the *low* and *mid* users, and instead, they are more active in the afternoon. From the weekday distribution, we notice a difference in the behavior of the *low* users and the *high* users: the *high* users check in mainly on working days, especially on Wednesdays and Thursdays, whereas for the *low* users, Saturday is the most popular day for checking in. Keeping track of visited places is one of the reasons why people use 4sq [18] so the least active users probably check in mainly when they visit new places, instead of trying to achieve mayorships. On Saturdays, people are typically free to explore new urban places which can explain the popularity of Saturday among the least active 4sq users. Finally, Sunday is overall the least checked-in day for both systems, which is not surprising given that shops and restaurants are mostly closed, except in touristic spots.

4.3 Is There a "Universal" Rank Distribution?

We now move to address fundamental questions related to place transitions and how they are captured by manual and automatic check-in systems. Noulas et al. [20] recently introduced a model for human mobility that characterizes the transitions people make in urban areas. The intuition behind the model is that the probability of a person visiting, e.g., restaurant X does not depend on the distance to X but rather on the number of other restaurants that are closer to X, i.e. the *number of intervening opportunities* (= *rank*). Formally the rank is defined as the number of places that are closer to the starting point of the transition than the destination of the transition. Furthermore, the model states that the distribution of ranks is universal across cities with different population densities, and it follows a power law for which an exponent of $\alpha = -0.88$ was found in [20].

We study whether this formula holds also for the MDC data and the Swiss 4sq data set we have. In order to make the two data sets comparable, we look only into the transitions that happen within the French speaking part of Switzerland since the MDC visits take place mainly there. As a rough approximation for this area, we consider only places whose longitude is between 6° E and 7° 30′ E (see Figure 2). We calculate the ranks of the transitions of a MDC user using only the user's own places, since if we would include the MDC places of all users, we would sometimes have the same places incorrectly appearing multiple times. The rank distributions for MDC and 4sq are shown in Figure 7.

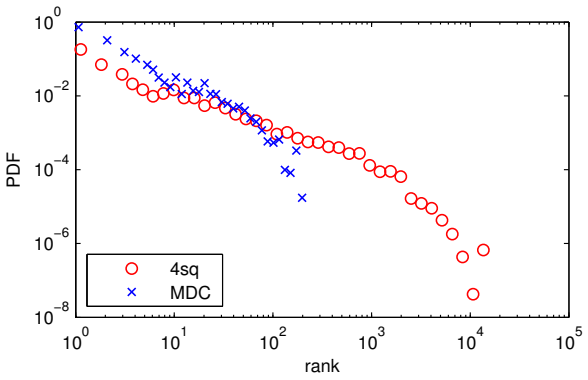

Figure 7: MDC and 4sq rank distributions which do not appear to follow the model proposed in [20].

The MDC distribution seems to have a steeper slope meaning that the MDC users are less likely to make transitions with a high rank. However, this is natural since a transition with rank 200 corresponds to a much longer travel distance for a MDC user than for a 4sq user. The reason for this is that the places of a single MDC user, while being "real" in the sense that they are the places that the user visits, actually are scattered much more sparsely than the places of all 4sq users.

We then redefine the rank of a MDC transition as the number of 4sq places that are closer to the starting point of the transition rather than the number of user's own MDC places. The modified rank distribution is shown in Figure 8.

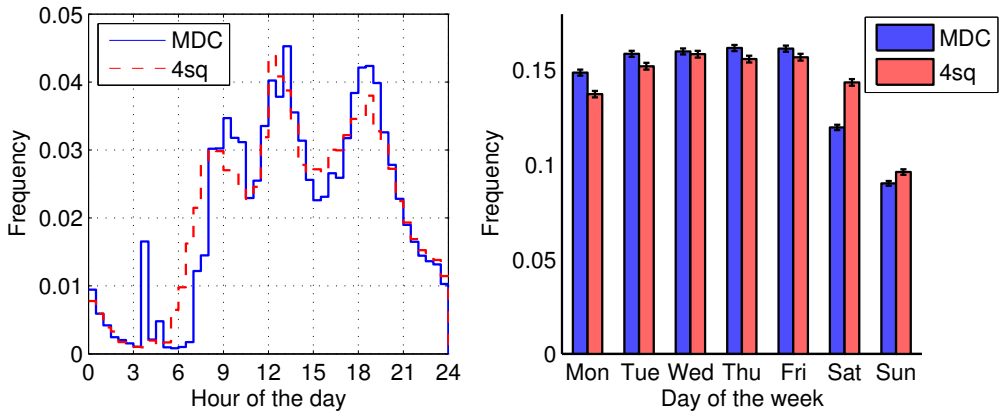

Figure 5: Distribution of visits over hours of the day (a) and days of the week with standard errors for the proportions (b).

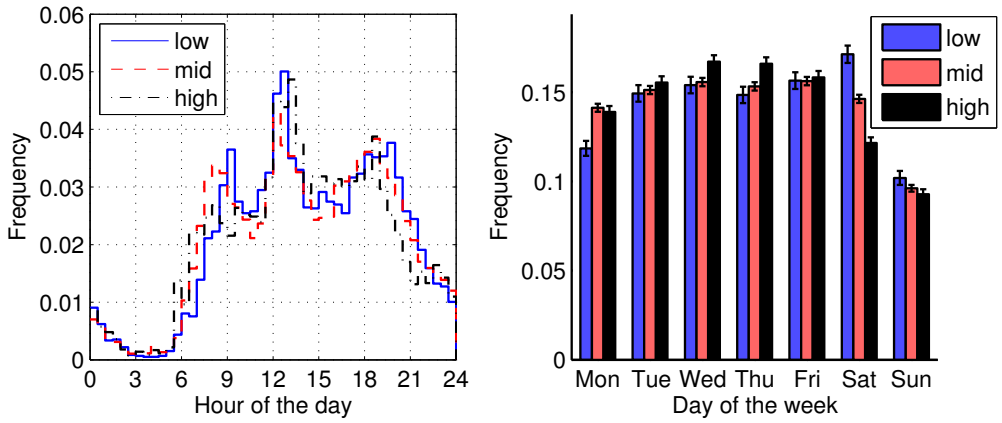

Figure 6: Visit time distributions for three 4sq user categories defined by their average check-in activity. low: users with $vpd < 1.5$; mid: users with $1.5 \leq vpd \leq 4.5$; high: users with $vpd \geq 4.5$. The weekly distribution shows also the standard errors for the proportions.

Now the two distributions follow each other very closely.

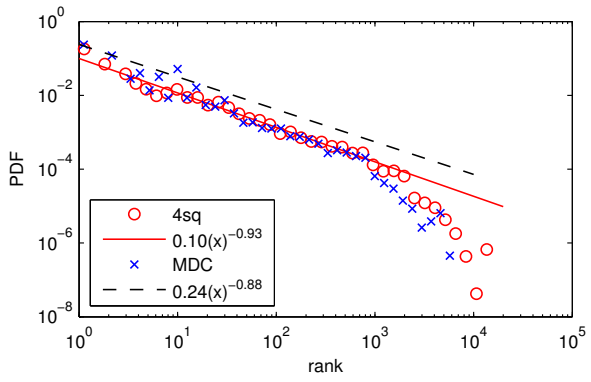

Figure 8: Rank distribution with the rank of a MDC transition defined as the number of 4sq places that are closer to the starting point of the transition than its end point. The dashed black line is the universal rank distribution given in [20].

Up to rank 10^3, they are also in accordance with the universal rank distribution estimated in [20] and shown with a dashed line, although with a slightly different exponent (-0.93) which has been estimated from the ranks below 10^3. However, from that on the distributions drop quickly to zero and do not follow a power law anymore. A natural explanation for this is that we consider only the transitions that happen within the western part of Switzerland. As a result, there cannot be many high rank transitions as there is not enough physical space to make them.

4.4 Next Place Predictability

We now address another basic issue, namely the predictability of places given automatic and manual check-ins. To measure the regularity of visit patterns, we test how well one can predict the next place based on the current place. Predicting is done with a *1st order Markov model*. Clearly, more sophisticated methods for prediction exist (e.g. see proceedings of the MDC Workshop [16]) but we are interested in observing trends that are clearly interpretable. We split the data into a training set, consisting of $\frac{2}{3}$ of the transitions in the beginning of the overall period for the users, and test set,

consisting of $\frac{1}{3}$ of the transitions in the end. A transition matrix for the Markov model is trained using the training set after which we evaluate the prediction accuracy on the test set. The accuracies are compared to a baseline classifier which predicts always the most common place. The results are shown in Table 3.

Table 3: Next place prediction accuracies with a 1st order Markov model and a baseline classifier which always predicts the most common place.

	Markov	baseline
MDC	47.4 %	34.7 %
4sq	19.5 %	12.6 %

These results confirm the finding that the visits in the 4sq data are less regular. This is what we expected since the visits are not necessarily consecutive as the users do not always check in when they visit a place. Furthermore, in the MDC data, the existence of visits to homes and workplaces increases the regularity of people's behavior. Nevertheless, some temporal patterns can be found in the 4sq data as well since the Markov model outperforms the baseline classifier. Furthermore, these results suggest that automatic check-ins seem a more promising approach as input to build prediction algorithms.

Very recently, a 1st order Markov model was applied to MDC in [9] and to 4sq data from some regions taken from all over the world in [8]. The reported accuracies are 43 % and 28 %, respectively. The accuracies differ from what we have shown since for the MDC data a different test set is used in [9] and for the 4sq data both the training set and the test set are different in [8]. In our case, both data samples come from the same country. Nevertheless, these reported accuracies also show that the MDC visit patterns are more predictable than the 4sq check-in patterns.

5. MATCHING AUTOMATIC AND MANUAL CHECKED-IN PLACES

So far we have investigated similarities and differences in temporal and transitional patterns for automatic and manual check-in systems. We now move to the problem of how to find correspondences across data sets that could be useful in their own sake, but that also highlight key aspects about the types of places found in each case. In this section, we compare the spatial distribution of places in the MDC and 4sq data sets and, more specifically, we want to study if it is possible to match the corresponding MDC and 4sq places.

5.1 Challenges in Finding the Matches

Let us take a look at MDC places and their nearby 4sq places. Figure 9 shows the distribution of distances from a MDC place to its closest 4sq place. We see that 50 % of the MDC places have at least one 4sq place within their radius of 100 meters, and 13 % have at least one 4sq place within 20 meters.

In general, the place matching problem is a difficult one. If there is only one 4sq place within 100 meters, it is trivial to match it to the MDC place, although obviously we cannot be sure if they are really the same place. However, in some

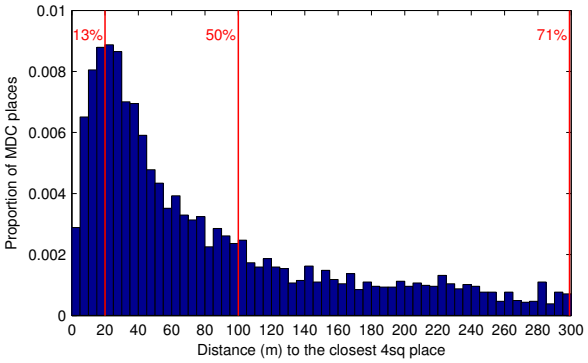

Figure 9: Distribution of distances from MDC places to their closest 4sq place. Vertical lines show the amount of probability mass on the left side.

cases there are more than one nearby 4sq place which we can see from Figure 10. Out of the 7281 MDC places, 33 % have at least two 4sq places within 100 meters; as an extreme case, one MDC place has 40 different 4sq places within 100 meters. In this kind of cases it may be difficult to determine which one of the neighboring 4sq places actually corresponds to the MDC place.

5.2 Nearest Neighbor Matching

Physical proximity is obviously a strong cue for matching places. We now consider a simple strategy where we always match the nearest 4sq place to each MDC place. We consider only the MDC places that have been labeled by the users (see Section 3.1) and that have at least one 4sq place within 100 meters. This corresponds to a total of 154 places.

To assess the accuracy of the method, we compare the MDC label to the category and title of the 4sq place and manually verify if they match. For instance, the MDC label *My workplace/school* and 4sq title/category pair *Batiment IN/College Engineering Building* would be considered a match. The mismatches, i.e., places whose function or meaning cannot be put in correspondence through manual inspection, are divided into two subcategories. The first subcategory *Fail1* refers to cases where the matched MDC and 4sq places are completely different, e.g., two nearby buildings. The second subcategory *Fail2* refers to cases where the MDC and 4sq places are inside the same building (e.g. two stores in the same mall) or the 4sq place is a subplace of the MDC place (e.g. 'university cafeteria' and 'university'). The motivation for this kind of distinction of mismatches is that while errors of the latter type (Fail2) are difficult to handle if we only use GPS data since GPS does not work well inside buildings, the errors of the first type (Fail1) may result from missing 4sq places. A 4sq place is missing if it has not been visited by the users in our data set during the data collection period of 6 months or if nobody has added it to 4sq, which is the case for a majority of the users' residences.

The results for the nearest neighbor place matching are shown in Table 4. We can see that a majority of the workplaces/schools (category C) are matched correctly. A typical example of a mismatched workplace/school is a MDC place that corresponds to the EPFL university while the nearest

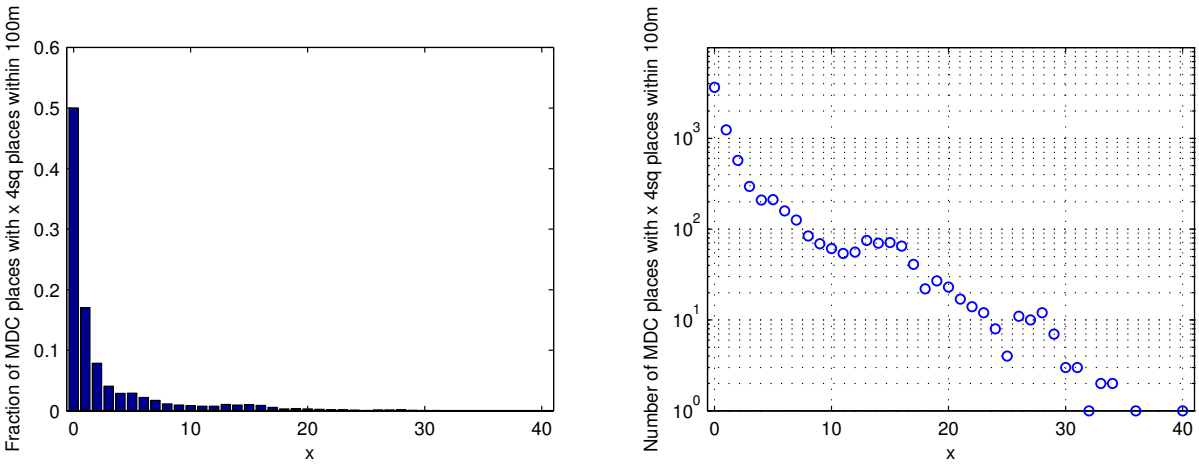

Figure 10: Left: probability distribution of x, where x is the number of 4sq places within 100-meter radius from a MDC place. Right: the same distribution with the actual numbers of MDC places falling into each bin (instead of probabilities) and a logarithmic scaling for y-axis.

4sq place corresponds to one of the EPFL cafeterias. Also *places for indoor sports* (category G) and *shops or shopping centers* (category I) have been matched accurately but for those categories we have less than 10 labeled MDC places. On the other hand, almost all of the *Homes* (category A and B) have been matched incorrectly, which is natural since 4sq users do not usually check in at their homes, and the likelihood of having MDC users who are at the same time 4sq users is very low given that the mobile platform used to record the MDC data (Nokia N95) is not very 4sq-friendly. If we exclude homes (categories A and B) and look at the places that are not isolated (i.e. have at least one 4sq place within 100 meters), we obtain a matching accuracy of 51 % for a total of 122 places.

The relation between the distance to the closest 4sq place and the probability of correctly matching a place are illustrated in Figure 11. The figure shows that if the closest 4sq place is within 40 meters, it is probably a match, whereas after 50 meters the matching probability starts to decrease. As a result, we can increase the precision of the matching by accepting only the 4sq places that are within 40 meters, whereas a threshold of 100 meters yields a higher recall. If we exclude homes (categories A and B) and look at the places that have at least one 4sq place within 40 meters, the matching accuracy of 67 % for a total of 60 places.

5.3 Improved Matching Methods

The previous subsection proposed a simple matching method based on the nearest 4sq place. To improve this approach, one could consider several nearest neighbors and make the choice between them based on other features than the distance. One alternative would be to look at the visit time distributions of the nearest neighbors and select the place whose distribution is closest to the visit time distribution of the MDC place, measured, e.g., by the Kullback-Leibler divergence. However, this would work only for a minority of places since most of the 4sq places have been visited only a few times and thus it is not possible to estimate the visit time distribution for them. The distribution of the number of visits per 4sq place is shown in Figure 12 and it seems to follow a power law.

Table 4: Place matching results and category descriptions. On the left, we have the MDC categories which are explained in the bottom table.

MDC Category	**Match**	Fail1	Fail2	Isolated	**Total**
A	2	21	0	63	86
B	0	9	0	37	46
C	42	11	22	29	104
D	7	8	3	4	22
E	0	2	1	6	9
F	4	4	0	17	25
G	4	1	0	9	14
H	0	5	0	6	11
I	5	3	0	9	17
J	0	0	0	5	5
Total	64	64	26	185	339

Type	Description
Match	Correct match.
Fail1	Matched MDC and 4sq places are completely different.
Fail2	MDC and 4sq places are different places but inside the same building.
Isolated	There are no 4sq places within 100 meters.

Category	Description
A	Home
B	Home of a friend, relative or colleague
C	My workplace/school
D	Location related to transportation (bus stop, metro stop, train station, parking lot, airport)
E	The workplace/school of a friend, relative or colleague
F	Place for outdoor sports (e.g. walking, hiking, skiing)
G	Place for indoor sports (e.g. gym)
H	Restaurant or bar
I	Shop or shopping center
J	Holiday resort or vacation spot

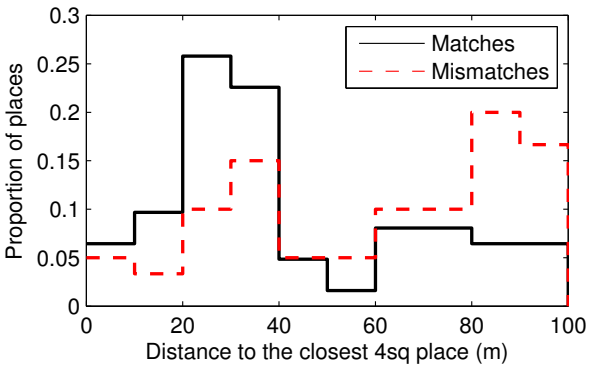

Figure 11: Distribution of distances to the closest 4sq place for the correctly matched MDC places and mismatched MDC places.

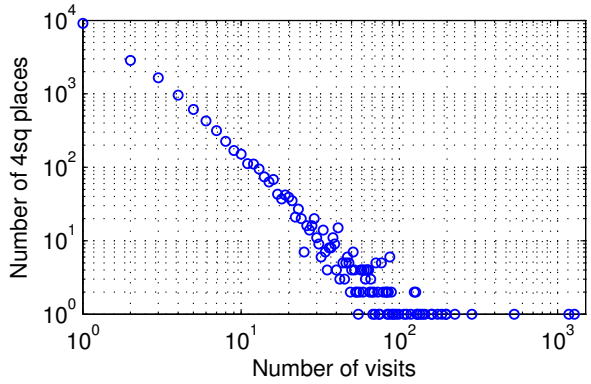

Figure 12: Distribution of the number of visits to 4sq places. On the y-axis we have absolute counts of 4sq places.

Another additional feature that could be used to select one of the nearest neighbors is the semantic labels that were used to assess the performance of the nearest neighbor method in the previous subsection. As explained in Section 3.1, a few MDC places have been given one of the labels shown previously in Table 4. For the rest of the places, we could estimate the labels following one of the approaches proposed for the MDC semantic place prediction task (see proceedings of the MDC Workshop [16]). Then we could select the 4sq place whose place category (see Section 3.2) corresponds to the semantic label of the MDC place. A limitation of this method is that we would first need to manually define which MDC label corresponds to which 4sq category.

6. FINAL DISCUSSION AND CONCLUSIONS

Motivated by their use in computational social science studies, we conducted a large-scale, country-level comparative analysis of data generated by two prototypical location check-in strategies, in which location is disclosed automatically (MDC) or manually (Foursquare). Partly due to the scarcity of public large-scale mobile sensor data, studies like ours have been missing in the literature. The two data sets have been collected in Switzerland, and so the mobility and social networking patterns reflected in the two data sets correspond to European trends.

First, we investigated fundamental aspects related to the type of check-in behavior revealed by the data sets. Our main findings are the following: (1) Overall, automatic checking-in results in more visits than manual checking-in. This difference is likely explained by natural limitations of human attention and interest, but also by privacy concerns of LBSN users regarding checking into private places [18]. At the same time, automatic and manual check-in levels are comparable when visiting new places, which is a basic incentive mechanism in LBSNs like Foursquare. This result suggests that participation in LBSNs through check-ins can reach similar levels to automatic sensing for specific settings and social situations. (2) The daily and weekly check-in activity patterns are similar for both systems, except for Saturdays – when manual check-ins are relatively more probable than automatic check-ins. This can be explained by the least active Foursquare users, who in general check-in only every now and then, but clearly have Saturday as their most active day. (3) A recently proposed rank distribution to describe human mobility, originally introduced in [20] and only validated on manual check-in data, also holds for automatic check-in data given a slight modification to the definition of rank. (4) The patterns described by automatic check-ins are in general more predictable, which is natural since manual check-ins overall occur less often and seem to respond more to novelty than to routine.

Second, we studied the key issue of finding correspondences between checked-in places across automatic and manual data sets using only location information. Given the resolution of the automatically discovered places in the MDC data, we showed that this can be a difficult problem in dense areas, where multiple nearby Foursquare places might correspond to a MDC place. In general, by using a simple nearest neighbor matching technique, an accuracy of 51 % was obtained for MDC places that had a Foursquare place within 100 meters and were not homes of the MDC users. This could be further improved up to 67 % by considering only places whose nearest neighbor was within 40 meters. Obviously, place matching could also be improved by using higher-level information (e.g. coming from maps); the validation of this idea would necessarily involve additional work.

Based on our study, we foresee two promising directions for future research. The first one is the development of a conceptual framework based on the above findings that could identify specific settings and users for whom automatic and manual check-ins could complement each other. For example, there might be users who could agree to have automatic check-ins in places they have previously visited, users who would like to authorize certain automatic check-ins, and users who would prefer to be in control of their check-ins all the time, for whom automation might not be useful. The second direction is the investigation of how data sets generated by different check-in strategies could be used to improve classification or prediction models learned from single check-in sources, e.g. via domain adaptation. Given our finding that the daily and weekly check-in activity patterns are relatively similar, this seems as a promising next step.

7. ACKNOWLEDGMENTS

This research was funded by the LS-CONTEXT project (Nokia Research Lausanne) and the HAI project (Swiss National Science Foundation).

8. REFERENCES

[1] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[2] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.

[3] Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring millions of footprints in location sharing services. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[4] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[5] H. Cramer, M. Rost, and L. Holmquist. Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 57–66. ACM, 2011.

[6] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[7] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):3, 2011.

[8] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[9] H. Gao, J. Tang, and H. Liu. Mobile location prediction in spatio-temporal context. In *Proceedings of the Mobile Data Challenge by Nokia Workshop in conjunction with International Conference on Pervasive Computing*, Newcastle, U.K., 2012.

[10] M. Gonzalez, C. Hidalgo, and A. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[11] J. Hightower, S. Consolvo, A. LaMarca, I. Smith, and J. Hughes. Learning and recognizing the places we go. In M. Beigl, S. Intille, J. Rekimoto, and H. Tokuda, editors, *UbiComp 2005: Ubiquitous Computing*, volume 3660 of *Lecture Notes in Computer Science*, pages 159–176. Springer Berlin / Heidelberg, 2005.

[12] J. Kang, W. Welbourne, B. Stewart, and G. Borriello. Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 110–118. ACM, 2004.

[13] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *Proceedings of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, Barcelona, Spain, April 2006. IEEE Computer Society Press.

[14] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proceedings of the 7th International Conference on Pervasive Services*, Berlin, Germany, 2010.

[15] A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, et al. Place lab: Device positioning using radio beacons in the wild. *Pervasive Computing*, pages 301–306, 2005.

[16] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Proceedings of the Mobile Data Challenge by Nokia Workshop, in conjunction with International Conference on Pervasive Computing*, Newcastle, U.K., 2012.

[17] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.

[18] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 2409–2418. ACM, 2011.

[19] R. Montoliu and D. Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 2010.

[20] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE*, 7:e37027, 05 2012.

[21] P. Nurmi and S. Bhattacharya. Identifying meaningful places: The non-parametric way. In *Proceedings of the 6th International Conference on Pervasive Computing*, pages 111–127, Sydney, Australia, 2008.

[22] K. Tang, J. Lin, J. Hong, D. Siewiorek, and N. Sadeh. Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 85–94. ACM, 2010.

[23] M. Ye, K. Janowicz, C. Mülligann, and W. Lee. What you are is when you are: the temporal dimension of feature types in location-based social networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 102–111. ACM, 2011.

[24] V. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World Wide Web*, pages 1029–1038. ACM, 2010.