Predicting human mobility with activity changes

Wei Huang[a], Songnian Li[a,*], Xintao Liu[a], Yifang Ban[b]

[a] *Department of Civil Engineering, Ryerson University, Toronto, Canada*
[b] *Division of Geoinformatics, Department of Urban Planning and Environment, Royal Institute of Technology, Stockholm, Sweden*

Abstract

Human mobility patterns can provide valuable information in understanding the impact of human behavioral regularities in urban systems, usually with a specific focus on traffic prediction, public health or urban planning. While existing studies on human movement have placed huge emphasis on spatial location to predict where people go next, the time dimension component is usually being treated with oversimplification or even being neglected. Time dimension is crucial to understanding and detecting human activity changes, which play a negative role in prediction and thus may affect the predictive accuracy. This study aims to predict human movement from a spatio-temporal perspective by taking into account the impact of activity changes. We analyze and define changes of human activity and propose an algorithm to detect such changes, based on which a Markov chain model is used to predict human movement. The Microsoft GeoLife dataset is used to test our methodology, and the data of two selected users is used to evaluate the performance of the prediction. We compare the predictive accuracy (R^2) derived from the data with and without implementing the activity change detection. The results show that the R^2 is improved from 0.295 to 0.762 for the user with obvious activity changes, and 0.965 to 0.971 for the users without obvious activity changes. The method proposed by this study improves the accuracy in analyzing and predicting human movement and lays a foundation for related urban studies.

Keywords: Activity change, Human mobility prediction, Spatial-temporal pattern, Markov chain, Spatio-temporal clustering

1. Introduction

Predicting human movement becomes viable with increasingly available human mobility data, which can be obtained through various channels (e.g., GPS-enabled smart and wearable devices, social networks, and even interviews and surveys). Numerous literatures have been dedicated to the identification of human mobility patterns and prediction of human movement from different fields based on this kind of mobility data. Brockmann et al. (2006) estimated individual displacement by analyzing circulations of bank notes in USA and proposed a scaling law for modeling human mobility patterns. They addressed that the jumping

[*] Corresponding author

Email addresses: wei1.huang@ryerson.ca (Wei Huang), snli@ryerson.ca (Songnian Li), xintao.liu@ryerson.ca (Xintao Liu), yifang.ban@abe.kth.se (Yifang Ban)

Preprint submitted to International Journal of Geographical Information Science *March 17, 2015*

size and waiting time in user-generated trajectories follow a scale-free distribution, where long travel distances and stay time are rare while short ones are common. Such patterns are claimed to be consistent at any reasonable spatio-temporal scale. Gonzalez et al. (2008) used phone calls or text messages and the location of the signal towers to reconstruct users' time-resolved trajectories. Their research found that people are intended to return to a few most frequently visited locations, which can be characterized by a single spatial probability distribution that indicates the dynamics behind the reproducible scaling patterns. By using similar data presented in Gonzalez et al. (2008), Song et al. (2010a,b) proposed two generic mechanisms, i.e., occasional exploration and preferential return that govern human trajectories from a microscopic perspective, and they further claimed that 93% of potential predictability in human mobility.

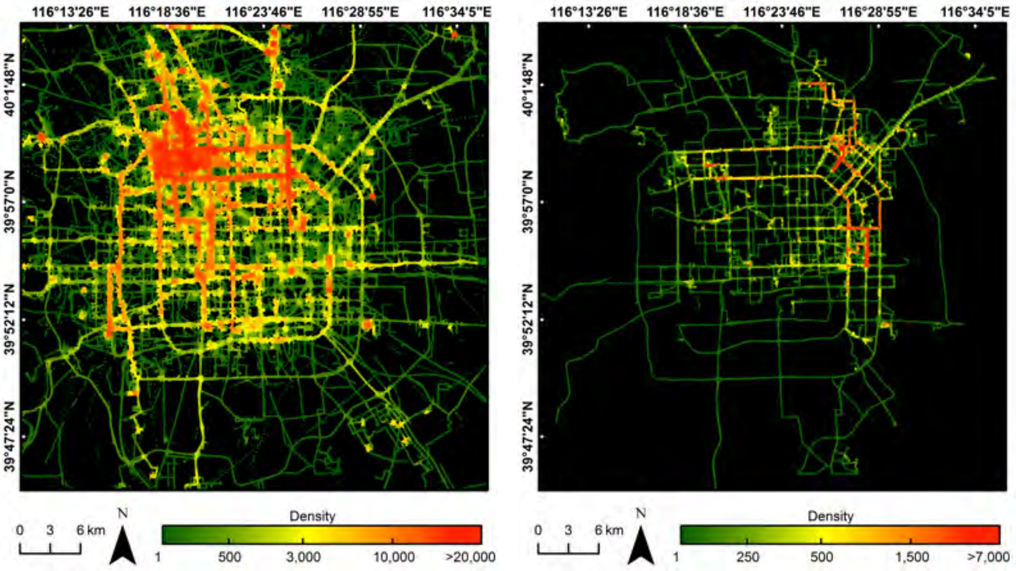
As more detailed and fine-grained user mobility data became available, a series of studies shed a deeper light on understanding and predicting human movement rather than just modeling general mobility patterns. Ashbrook and Starner (2003) suggested that the most significant places where a user stays should be those the user spends some of his/her time (i.e., a user stays somewhere for a certain amount of time to do something). They extracted user activities with a given time threshold (e.g., 20 minutes) and then clustered the spatial location of activities into groups, based on which a graph was created and a Markov model was applied to predict. Similarly, Liao (2006) extracted and clustered user activity locations and proposed sophisticated Markov models to recognize and infer human activities. Their method is capable of learning specific motion patterns of transportation routines to detect user errors. Besides, Asahara et al. (2011) proposed a state-space modeling method to predict pedestrian mobility using a mixed Markov Chain model (MMM). Gambs et al. (2012) developed a mobility model called the Mobility Markov Chain (MMC) to predict the next place to travel. Mathew et al. (2012) proposed a predictive model based on Hidden Markov model (HMM), which uses a timestamp (i.e., 7 a.m. to 7 p.m. of weekday, 7 p.m. to 7 a.m. of weekday, and weekend) to associate the places in sequences. Coincidentally, Zheng et al. (2011, 2009) defined the concept of stay point to extract user activities based on two randomly and empirically determined time and distance thresholds (e.g., 20 minutes and 200 meters), and created a hierarchical structure to infer the top interested user locations and travel sequences. Despite the above efforts on human movement prediction, time dimension is usually being treated with oversimplification or even being neglected. To our knowledge, the study by Mathew et al. (2012) that does consider time dimension in prediction roughly is the only one that determines three time slots in a manual way, but calls for more proper ways to determine time dimension in its recommendations for future work.

In order to better predict human mobility in a more precise manner, time dimension should be considered to detect human activity changes because such historical human mobility changes (home change, school transfer or job change) have negative impacts on movement prediction. For example, Song et al. (2006) introduced an aging mechanism to reduce the negative impact of changed locations toward improving the prediction performance. A multiplicative factor was proposed to intervene in the process of sample training, and to reduce the contribution of old samples where changed locations are involved. Etter et al. (2013) proposed an algorithm that is able to detect home changes. In these two studies, the datasets with semantic information were used to detect home changes. The authors assumed that each person has a single home, and once another home is subsequently detected during the model training, the detected home is flagged as the new home. Although these studies investigated the historical changed locations (actually they attempted to detect home changes only), they can only deal with datasets with semantic information, which is not applicable for most user-generated trajectory datasets such as the Microsoft GeoLife dataset. Inspired by the related studies, we develop a methodology that can detect human activity changes from a non-semantics human trajectory dataset (Microsoft GeoLife dataset), based on which we apply a Markov chain model to predict human movement.

The paper is organized as follows. Section 2 introduces the dataset used for this study. A methodology including activity extraction, activity clustering, activity change detection, and human movement predicting is developed in Section 3, followed by validation of the methodology. Conclusions and some future works are presented at the end of the paper.

2. Dataset

The dataset, obtained from the Microsoft GeoLife project, contains 17,621 GPS trajectories that were collected by 182 users over 6 years (from April 2007 to August 2012) using different wearable GPS receivers and cell phones (source: http://research.microsoft.com/en-us/projects/geolife/). These trajectories contain 25,298,541 GPS points, covering a total distance of about 1.2 million kilometers and a total duration of more than 48,000 hours (c.f., Zheng et al. (2009) for more details). These trajectories recorded a broad range of users' outdoor movements, not only daily routines such as staying at home and working, but also leisure activities such as shopping, sightseeing, dining, hiking and cycling. Figure 1 illustrates the trajectories of all users as well as a specific user #153 in Beijing, China (lines in color). The color shows the density of the GPS points determined based on a 200 m by 200 m window.



(a) All users

(b) User #153

Figure 1: Trajectories of all users (a) and a specific user #153 (b) in Beijing, China

3. Methodology

We propose a four-step methodology (Figure 2) which consists of 1) extracting human activities from raw GPS trajectories, 2) clustering activities in terms of spatio-temporal dimension using density-based spatial clustering of applications with noise (DBSCAN) method (Ester et al., 1996), 3) detecting two types of human activity changes based on the identified activity clusters, and 4) predicting human movement based on Markov chain model with consideration of activity changes.

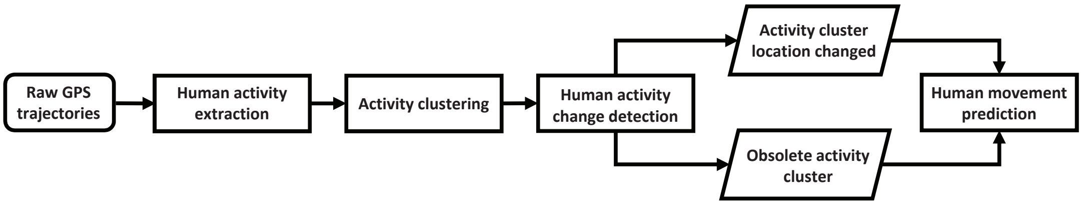

Raw GPS trajectories
Human activity extraction
Activity clustering
Human activity change detection
Activity cluster location changed
Obsolete activity cluster
Human movement prediction

Figure 2: The methodology flowchart

3.1. Extraction of human activity

A human activity can be defined as a meaningful location of a user where he/she spends some time (Ashbrook and Starner, 2003). We use the method proposed by Zheng et al. (2011, 2009) to extract human activities from raw GPS trajectories based on a time threshold of 20 minutes and a distance threshold of 200 meters. This means that an activity is defined if a user stays for at least 20 minutes in a geographic region with a diameter of 200 meters.

Definition 1. *A* ***single activity*** *is an activity (point) extracted from raw GPS trajectories.*

In total, 38,520 single activities are extracted from 182 users over 6 years (see Table 1). Some users do not have any activities, while 7 users have more than 1,000 activities. Single activities of all users are plotted in Figure 3. According to Figure 3, it is obvious that most of single activities are located in the north-west part of Beijing, China, where a few universities and research institutes (e.g., Peking University, Microsoft Research Asia, etc.) are situated. Therefore, we can infer that most of users come from these academic institutions.

3.2. Activity clustering

Normally, GPS receivers are not able to record the identical coordinates at the same location due to their positioning error, even though users precisely stay at the same point of a place. More importantly, a series of activities may occur at different times, i.e., their arrival times are different (e.g., the time of arriving home can be either in the noon or in the evening). Therefore, it is necessary to perform spatio-temporal clustering to classify single activities considering temporal aspect and further filter single activities in terms of spatial dimension, which is the key to detect human activity changes and build a predictive model in our work later on.

Table 1: The number of single activities of each user extracted from the Microsoft GeoLife dataset

UID	A. A.	UID	A. A.	UID	A. A.	UID	A. A.	UID	A. A.	UID	A. A.	UID	A. A.	UID	A. A
153	2951	140	424	20	176	134	100	95	48	154	25	76	12	27	4
128	2026	52	371	15	170	155	98	94	48	114	24	54	12	148	3
4	1487	92	349	12	164	40	93	66	48	93	24	156	11	72	3
3	1233	104	341	179	151	74	92	57	48	79	24	127	11	149	2
17	1122	34	323	13	151	147	85	23	48	176	21	77	11	137	2
163	1113	37	319	9	151	83	78	141	47	106	21	80	10	116	2
68	1050	112	296	71	148	131	75	55	44	164	19	143	9	31	2
30	993	36	292	43	141	33	73	102	42	45	19	100	9	21	2
167	783	42	290	73	138	29	66	169	41	99	19	124	8	177	1
85	751	24	251	174	135	122	65	56	40	59	18	121	8	172	1
144	680	5	249	125	132	81	65	181	38	161	16	63	8	133	1
41	663	67	232	10	129	130	64	58	38	98	16	45	8	60	1
22	610	96	225	91	128	89	64	46	37	61	16	117	7	48	0
25	607	65	214	28	126	6	63	138	36	129	15	87	7	49	0
39	591	115	210	78	124	159	62	157	34	158	14	70	7	66	0
0	562	82	195	16	124	103	59	113	33	146	14	53	7	118	0
62	548	44	186	168	123	165	57	105	31	108	14	175	6	120	0
14	511	11	186	18	117	32	57	135	28	47	14	166	6	123	0
35	503	1	185	19	112	64	56	111	28	152	13	162	6	132	0
126	490	7	184	51	109	97	53	139	27	90	13	109	6	160	0
38	451	142	182	26	107	88	53	136	26	173	12	170	4	180	0
84	434	119	179	8	106	110	50	75	26	171	12	151	4		
2	433	50	177	20	176	150	48	69	26	86	12	107	4		

Note: UID represents User ID and A. A. represents Activity Amount.

Definition 2. *An **activity cluster** is a group of single activities happened at the same location in a specific time duration.*

We use specific time and space thresholds to define search radius and set 1 as the minimum number of points required to form a dense region. The DBSCAN is applied to perform the required clustering analysis. If the "distance" in terms of time and space amongst points is within the specified thresholds, these points are grouped as a single cluster (Figure 4). The accuracy of GPS receivers is used as the space threshold since the deviation between two coordinates of the same location is mainly caused by the error of GPS receivers. A threshold of 15 meters, therefore, is set as the space threshold in this paper, since it is the average error of the GPS receivers used in the GeoLife project. Since there is no universal number that could be used to describe each user's time characteristics, we develop a data-driven method to determine this threshold for different users.

The distribution of time interval between a single activity and the subsequent single activity varies with different individuals but overall follows a Heavy-tailed distribution (Figure 5). According to Jiang et al. (2013), if a series of variables follow a Heavy-tailed distribution, the mean can separate all the variables into a high percentage of small ones and a low percentage of large ones. Thus, the mean can represent the characteristics of the major population. Based on this theory, the mean of time intervals carries some underlying meanings: similar and continuous single activities happen within a certain period of time. In other words, single activities happened around a time within the mean of time intervals have similar characteristics in terms of time aspect. As a result, the mean of time intervals

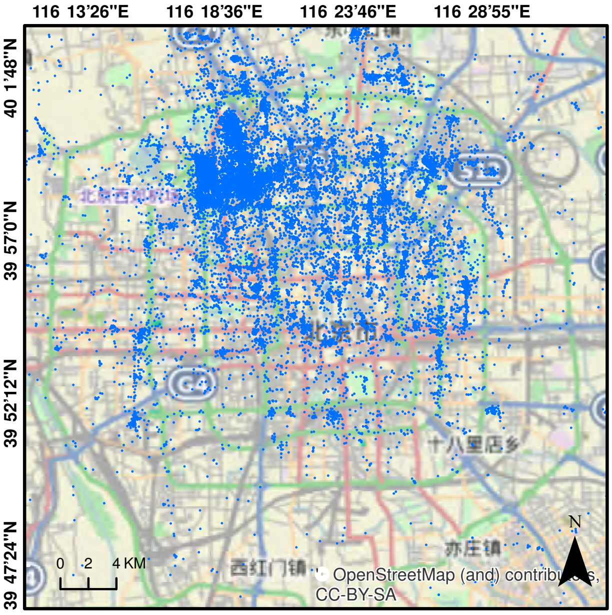


Figure 3: The geographic distribution of all users' single activities in Beijing, China

is used as the time threshold to cluster the corresponding user's single activities. The two thresholds are thus identified: 15 meters (as mentioned above) as the space threshold and the mean of time intervals as the time threshold. Every single user may have different time threshold, which entirely depends on the user's temporal pattern of the extracted single activities.

If a single activity of one activity cluster and a single activity of another activity cluster are within the space threshold (15 meters), these activity clusters are considered as located in the same place, which is represented by the centroid of the single activities from these different activity clusters.

3.3. Human activity change detection

Human activity changes play a negative role in predicting human movement because a new location of an activity cluster recently emerged in the training sample has lower transition probability than the previous location within a short period. In other words, the previous location of that specific activity cluster still has a higher transition probability than that of the new location if the change occurred close to the end of the sample data for training predictive model. For instance, a person recently moves to a new apartment A_1 from the previous apartment A_2. If the next possible place where this person is heading to is her/his home, the most possible location of her/his home is still the location of A_2 since it has a higher transition probability than the location of the apartment A_1. Actually, the correct predicted location of the home should be the location of the apartment A_1 regardless how high the transition probability of the apartment A_2 is. Hence, we aim to model such activity change by providing a definition of activity change and a change detection method.

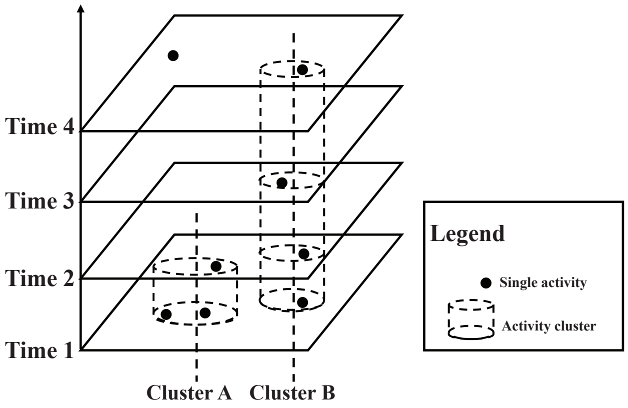
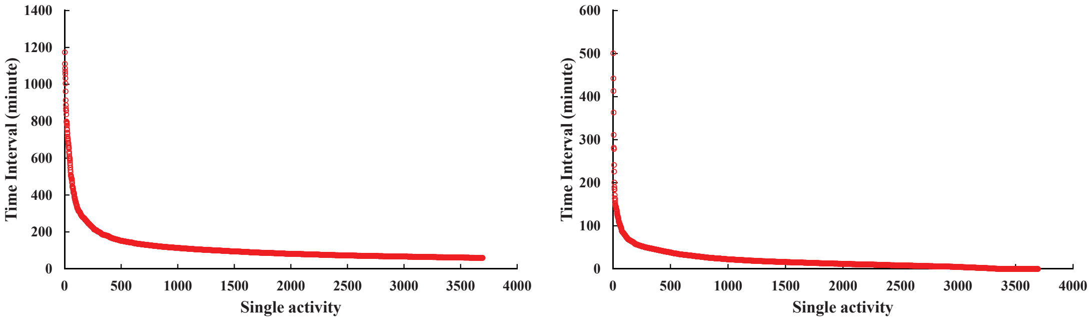


Figure 4: Spatio-temporal clusters



(a) All users

(b) A single user

Figure 5: Distribution of time intervals between single activities

Definition 3. *A human activity change occurs when i)* ***activity cluster location changed****; or ii)* ***activity cluster obsoleted****. Some human activity changes, such as job change (the location of office changed) and school transfer (the location of campus changed), are related to the change of activity cluster location. Other human activity changes, such as graduation (the activity of studying ended) and retirement (the activity of working terminated), can be seen as activity cluster obsoleted.*

In order to detect how activity cluster location changes, the two methods proposed by Song et al. (2006) and Etter et al. (2013) both can detect a person's home change through analyzing if the location of a home coming after the previous one in the training samples is the same as the location of the previous home. If the answer is no, the home of this person is changed. This kind of method is based on an assumption that every person has only one single home. When two homes are found in different locations, a home activity change is identified. Based on that, we can make an assumption that one person can only have one type of activity during specific time duration. For example, one person generally only has one single working place, school, or gym in a given time duration. Furthermore, since there is no additional information in GPS dataset indicating the type of place visited (e.g., home, office or school), it is impossible to directly obtain the name of the activities extracted from GPS trajectories. As a result, we calculate similarity between activity clusters and examine if similar activity clusters occurred sequentially in terms of date. If they are similar and occur sequentially, they are considered as changed activity clusters. The advantage of the proposed method is that we can detect the changes of certain activities that are difficult to infer their semantics. Two parameters, arrival time of activity cluster (A_t) and stay duration of activity cluster (S_d), are used to define the similarity of activity clusters (Figure 6a). If any two activity clusters occur at similar time and their stay durations do not have significant difference, such activity clusters can be considered as similar, representing a certain type of activity. To quantitatively define the similarity among activity clusters based on the work of Chen et al. (2011), the overlapping area between two rectangles enclosed by arrival time and stay duration of any two activity clusters is used to measure the similarity. More specifically, the equation proposed by Tversky (1977) based on a similarity theory in psychology is utilized to quantitatively define the similarity between two activity clusters, which can be expressed by:

$$D = \frac{c}{a + b - c} \tag{1}$$

where a and b are the area of the two rectangles enclosed by arrival time and stay duration, respectively (Figure 6b); and c is the overlapping area of a and b. If two activity clusters have same arrival time and stay duration (i.e., $a = b$), the similar measure D equals to 1. In other words, the higher level the similarity between two activity clusters, the closer the value of D is to 1.

In this study, two activity clusters can be considered as similar activity clusters if $D > d$, where d is a threshold and can be computed by:

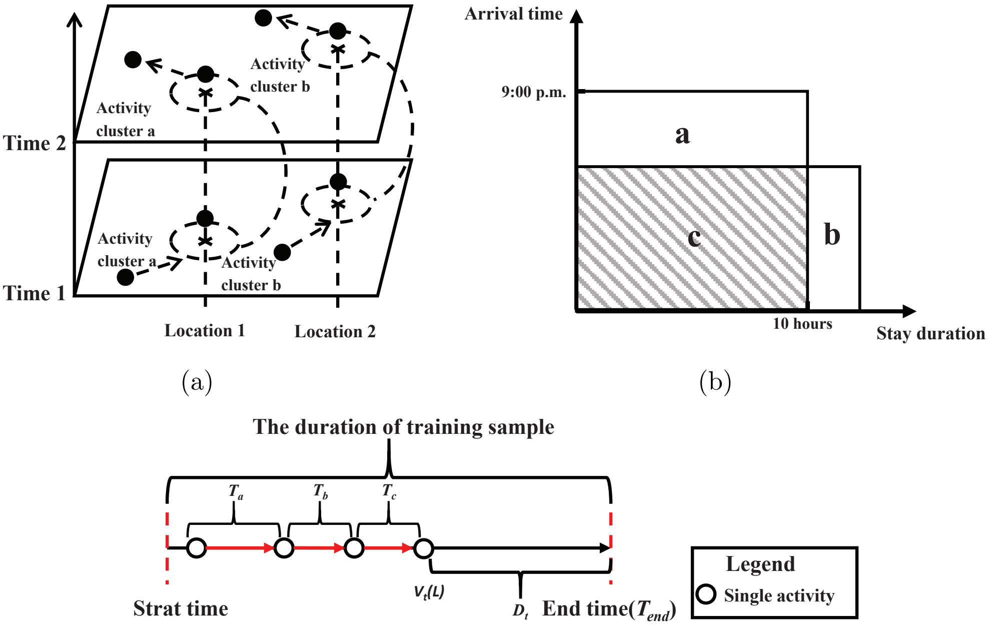
$$d = \frac{c}{\sum_{i=1}^{n} a_i - (n-1)c} \tag{2}$$

where a_i refers to the area of the rectangle enclosed by arrival time and stay duration of a single activity in an activity cluster; n refers to the number of single activities in that activity cluster; c denotes the overlapping area of all the activities in that activity cluster. In this case, d represents the similarity of all activities in the clusters, thus it can be used to determine if another activity cluster have the same similarity.

On the other hand, to detect obsolete activity clusters, we compare the average time interval of single activities happening in an activity cluster with the time difference between the last time a single activity happened in this activity cluster and the end time of training sample (see Figure 6c). The average time interval of single activities in an activity cluster can be defined as follow:

$$T_{int} = \frac{\sum T_a}{n - 1} \tag{3}$$

where T_a refers to the time interval between two neighboring single activities in terms of happening time in an activity cluster; n denotes the number of single activities in an activity cluster. The time difference can be defined as follow:



(a)

(b)

(c)

Figure 6: Similar activity clusters (Activity cluster a is similar with Activity cluster b). The arrival time of these activities is $Time1$, and the duration the person stays on activity a and activity b is $Time2 - Time1$ (a); similarity measure (b); and obsolete activity cluster (c)

$$D_t = T_{end} - V_t(L) \tag{4}$$

where T_{end} refers to the end time of the training sample, $V_t(L)$ is the time of the single activity in the activity cluster happened last time. If $D_t > T_{int}$, then this type of activities is considered as obsolete activity cluster, which means this activity cluster will not be taken into account for prediction. For instance, a dataset from January to June is used to train the predictive model. If we find that single activities in one activity cluster occurred every day, i.e., the average time interval is about one day, but did not continue within the last month before the end of June, this activity cluster should be considered as obsolete and should not be used in prediction.

Based on the above discussion, we propose an algorithm to detect the activity change. The pseudo-code of the process is shown in Algorithm 1.

3.4. Prediction of human movement

Markov chain is a well-known model for analyzing sequential data, which can be defined as a sequence of random variables $X_1, X_2, X_3, \ldots, X_n$ with the Markov property and, given the present state, the future and past states are independent (Markov, 1971). Formally, a Markov chain can be described as: $P_r(X_n + 1 = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P_r(X_{n+1} = x_{n+1} | X_n = x_n)$, where the possible values of X_i form a countable set S is called the state space of the chain. Markov chain is often described by a directed graph, of which the edges are labeled by the probability of going from one state to another state. Figure 7

Algorithm 1: Activity change detection algorithm

```
input  : all activity clusters of a single user
output: updated activity clusters of that single user
begin
    detect similar activity cluster(s) using $Eqs.$(1)&(2) for each activity cluster;
    if there is no similar activity cluster found for some activity clusters then
        calculate $T_{int}$ using $Eq.$(3) and $D_t$ using $Eq.$(4) for these activity clusters;
        compare $D_t$ and $T_{int}$ for these activity clusters;
        remove those activity clusters whose $D_t > T_{int}$;
    else
        use the location of the newest activity cluster as the location of all its similar
        activity clusters;
        update the activity clusters;
    end
end
```

illustrates an example of a classic Markov chain with three states describing by a directed graph. From X_1 to X_2, X_2 to X_3, X_3 to X_1, the probability is p_1, p_2, and p_3, respectively.

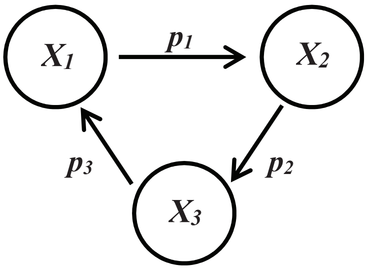


Figure 7: An example of Markov chain

We aim to predict the human movement at a specific time and a specific place, which depends purely on the current location and time without considering the preceding states. It inherits the typical Markov property, we therefore develop a human movement prediction model based on the concept of Markov chain.

First, all activity clusters are preprocessed by activity change detection algorithm (Algorithm 1), which enables changed activity clusters to be preprocessed. These preprocessed activity clusters are then used to train Markov chains. The activity clusters are utilized to set up the states of Markov chains. The transition probability between two states denotes the probability of a specific user traveled from one place (the centroid of an activity) to another at a specific time. This transition probability from state I to state J at time t is computed by:

$$P^t(J|I) = \frac{\sum T_{ji}}{\sum_{k=m}^{j} \sum T_{ki}} \tag{5}$$

where T_{ji} denotes a transition from state I to state J; m refers to the state connected to state i. Each activity cluster is associated with a specific Markov chain. In this case, the Markov chains related to the same place may be different if more than one cluster exists at the same location. For example, in Figure 8, at location C, three activity clusters at different time, Time a, b and c, are found. Thus, there exist three Markov chains associated to location C. At Time a, the transition probability from location C to B and D are 90% and 10%, respectively, and 10% and 90%, respectively at Time c. At Time b, there is a probability of 80% traveling to location A, but no transitions to location A exist at the other two times.

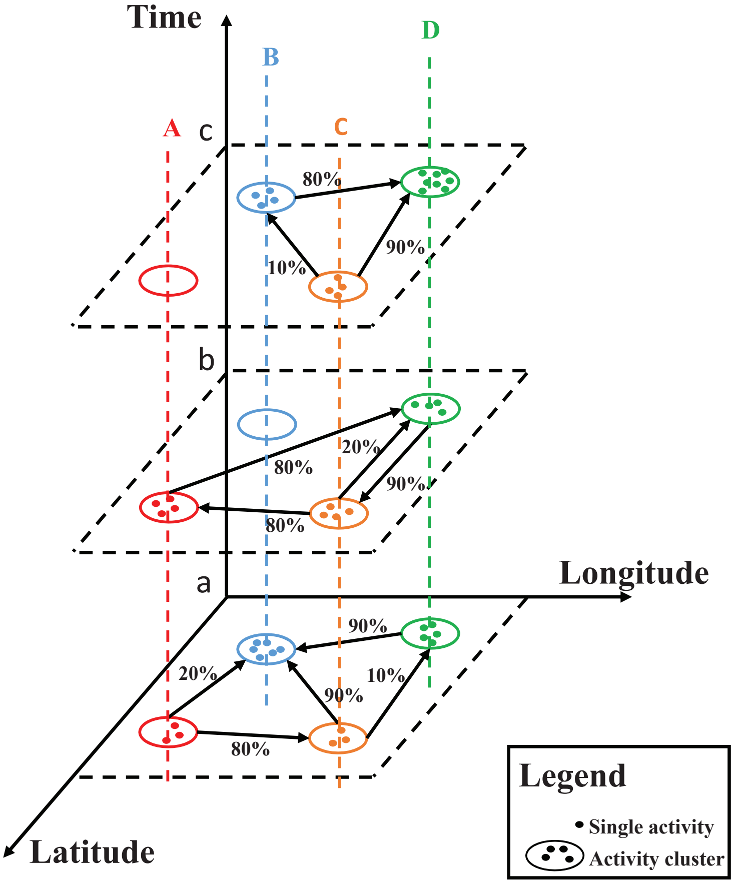


Figure 8: Markov chains at different times

As a result, the location most likely to be visited from one location varies at different times. Given a activity cluster set $S = (s_1, s_2, s_3, \ldots, s_n)$, the location most likely to be visited from location M can be defined as:

$$L(t) = \begin{cases} argmax(p^{t_1}(J|M)) \\ argmax(p^{t_2}(J|M)) \\ \ldots \\ argmax(p^{t_n}(J|M)) \end{cases} \tag{6}$$

Where $J \in S$, t_n denotes a specific time when some transitions are generated from location M.

Notice that the time t_n may actually be a very short time period since the time of each single activity occurred in one activity cluster may have difference in minutes. For instance, the time arrived home may be 10 to 15 minutes before or after 8:00 p.m. In this study, we treat this very short time period as a time point to simplify the description of the predictive model.

4. Results and validation

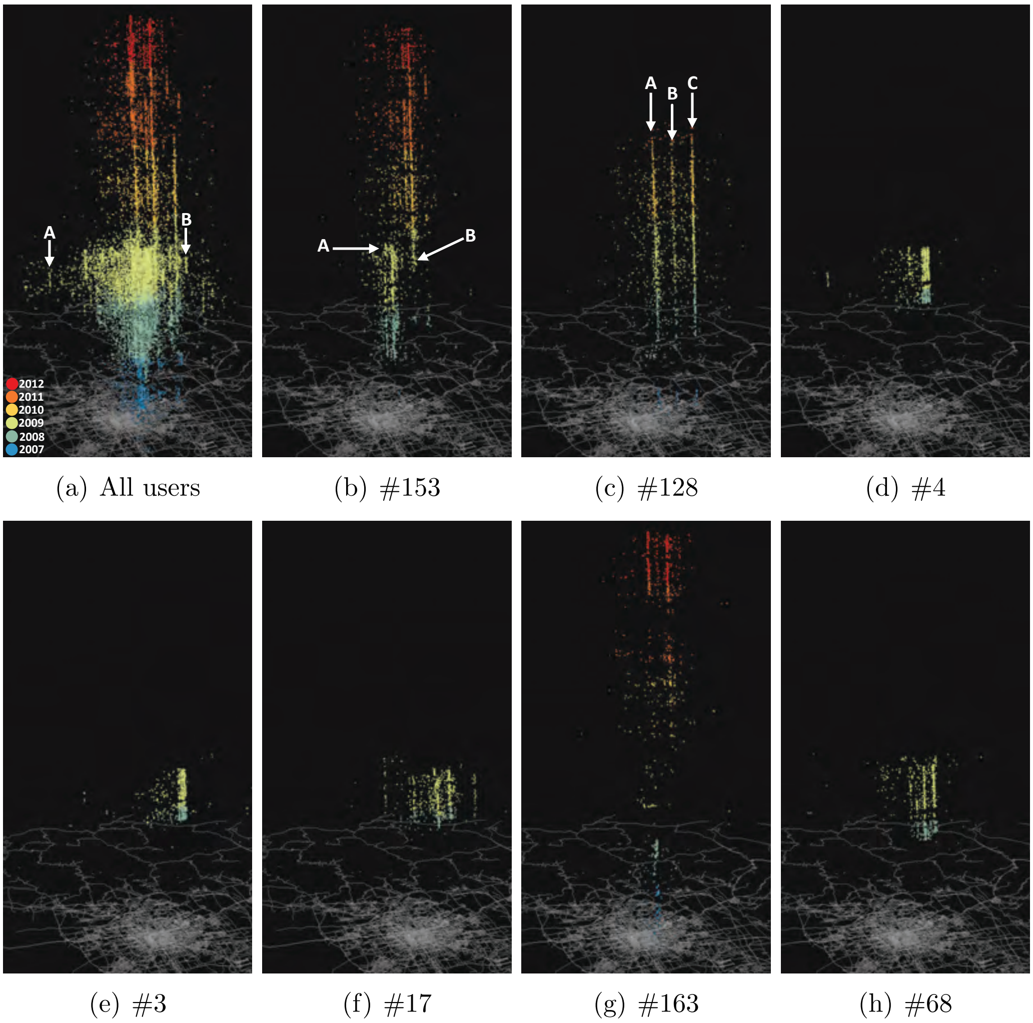
In this section, we first visually explore the extracted single activities on a 3D Geographical Information System (GIS) platform. We analyze the temporal patterns between two types of time attributes: arrival time and occurring duration of single activities. Based on that, we select two top users (#153 and #128) for experimental testing to evaluate the performance of the predictive model. The selection of these two users is mainly because they have two different typical mobility patterns. User #153 has obvious activity change pattern, while user #128 has three identical activities without changes. Furthermore, the spatio-temporal distribution of their single activities is more diverse than other users.

4.1. Human activity analysis

To capture the spatio-temporal regularities of human activities towards an understanding of human mobility pattern, we visualize all users' single activities and top users whose single activities are over a thousand in space and time dimensions (Figure 9). The height of the points in Figure 9 refers to time of day when the activity happened in minutes in each year, with the red representing the most recent year 2012, and the blue representing the earliest year 2007. Therefore, the total height in Figure 9 is minutes in 24 hours multiplied by the number of years. The horizontal distribution of the points is based on longitude and latitude coordinates, which are plotted on top of the major roads in the Beijing (shown as grey lines). To better explore the movement patterns and activity changes, we first define a location as a frequently and continuously revisited place as presented in a shape of vertical line in Figure 9. It is easy to note the obvious activity locations such as A and B for all users, A and B for user #153, and A, B and C for user #128. It should be noted that such locations of single activities found for both all users and individual users are aligned with the concept of spatio-temporal clusters.

For the single activities of all users, the geographic region of the single activities of all users (Figure 9a) in 2008 and 2009 is the most widespread and has higher density than any other years. We believe that such a phenomenon is mainly due to the 2008 Summer Olympics Games held in Beijing, China; the Olympic venues distributed around Beijing were visited more frequently in 2008 (the Olympics year) and 2009 (more tourism for the Olympics venues). Single activities for all users at location A and B just occurred for a short period during the year 2009. Such single activities are from different users, indicating some users temporally appeared in such locations (e.g., specific Olympics venues), and then they did not revisit again.

Those single activities associated with individual users (e.g., A and B for user #153 in Figure 9b) could carry some semantic information (such as home and working place).



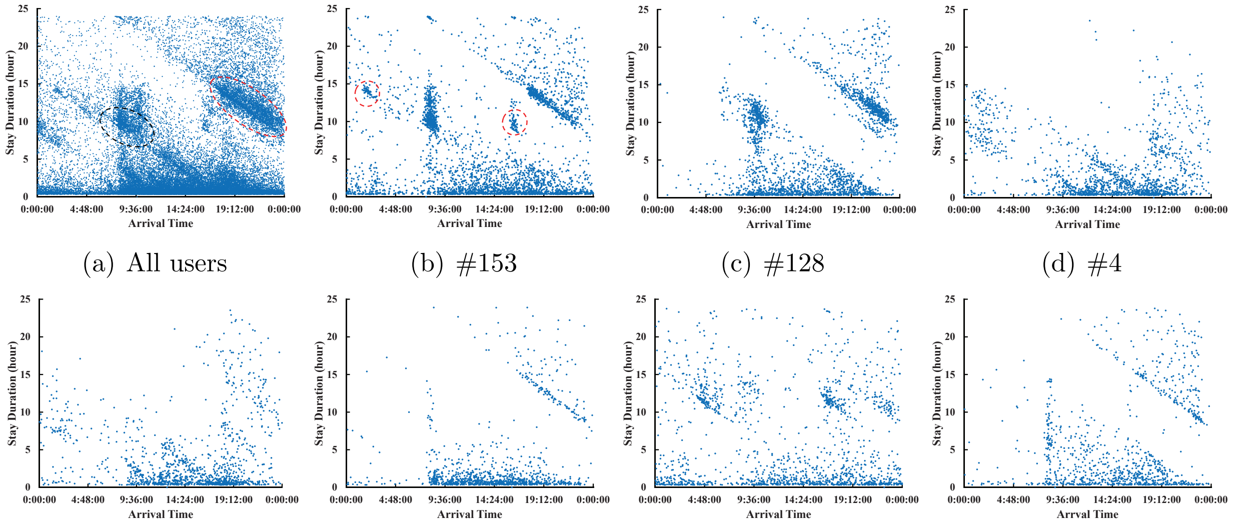
(a) All users (b) #153 (c) #128 (d) #4

(e) #3 (f) #17 (g) #163 (h) #68

Figure 9: Spatio-temporal distribution of single activities of all users and top users with more then 1,000 single activities. (a) All users. (b) User #153. (c) User #128. (d) User #4 (e) User #3. (f) User #17 (g) User #163. (h) User #68

Nevertheless, these fixed locations could change over time. As shown in Figure 9b, the activity pattern of user #153 demonstrates a pattern of activity change according to the **Definition 3**, i.e., the activities changed from location A to location B in 2009. In contrast, the activity pattern of user #128 (Figure 9c) demonstrates a type of stable pattern, where he/she mostly stayed at location A, B and C. In addition, there are some similarities of human activity pattern if we compare the single activity distribution of different users. For example, user #4 has a similar activity pattern with user #3, as shown in Figure 9d and 9e, since their single activity distributions in terms of temporal dimension and spatial dimension do not show significant visual differences.

We also plot the temporal patterns based on stay duration versus arrival time for the single activities of all users and top users who have over a thousand single activities in Figure 10. The vertical axis represents the stay duration of single activities in hours and the horizontal axis refers to the arrival time of single activities in Beijing time. Overall, all users



(a) All users

(b) #153

(c) #128

(d) #4

(e) #3

(f) #17

(g) #163

(h) #68

Figure 10: The temporal patterns of stay duration and arrival time: (a) All users; (b) User #153; (c) User #128; (d) User #4; (e) User #3; (f) User #17; (g) User #163; (h) User #68

have two significant stay durations of approximate 9 to 14 hours in total. One is related to the arrival time from around 8:00 a.m. to 9:30 a.m., marked by the black circle in Figure 10a. The other happened at the arrival time from around 6:00 p.m. to around 12:00 a.m., marked by the red circle in Figure 10a. A significant number of users have single activities that last for 9 to 14 hours in terms of stay duration, and arrival time spreads in between 6:00 p.m. to around 10:00 p.m. As a specific example, user #153 has another two different arrival times lasted for approximate 10 hours and 14 hours, marked by the red circles in Figure 10b.

Through analyzing the temporal patterns of single activities, some locations, such as home, office and gym, could be identified according to the pre-existing human behavioral regularities. For instance, in general, home and office/school are the two distinct locations where people spend longer hours (in average 8 - 10 hours) than that of the other activities, while the start time of these two types of places is very different. Normally, the time of arriving home is in the evening or at night since people usually go home directly after work/school or leisure activities, such as dinning or fitness training in gyms. In contrast, the arrival time of office/school is expected to occur in the morning. Thus, the distribution of single activities in Figure 10 reflects the reality. There are two time windows for arrivals, morning and evening to night, with stay duration of approximately 9 - 14 hours.

Since most activities are located in the north-west part of Beijing where several universities (e.g., Peking University, Tsinghua University, etc.) and research institutes (e.g., Microsoft Research Asia, Chinese Academy of Science, etc.) are located, we believe that users in the GeoLife project worked in these universities and research institutes. For university students, faculty and staff, they travel in between different classrooms, dine in the cafeterias, and have recreations, such as soccer and basketball, at outdoor playground. Consequently, Figure 10 reflects high density of activities during time durations of 1 hour (for meals) and 2 hours (for class or recreations after classes). To summarize, we found some common characteristics by visually analyzing these activities such as the location of certain changed activities. The semantics of some places could be inferred through analyzing the temporal patterns of activities.

4.2. Activity clusters exploration

All users' single activities are clustered, from which we select top 5 sample clusters of two typical users (User #153 and User #128) based on the heavy-tail distribution of the number of single activities in each activity clusters for further analysis. All the single activities of these selected clusters are projected from WGS84 to UTM Zone 50N for visualizing them in 3D. Figure 11 shows the distribution of the selected activity clusters of User #153 and the activity distribution in each of these cluster. In the figure, the x-axis refers to the east direction in meter, whereas the y-axis refers to the north direction in meter. The specific (x, y) coordinates represent where a sing activity occurred and the z-axis refers to the time of day when the single activity happened.

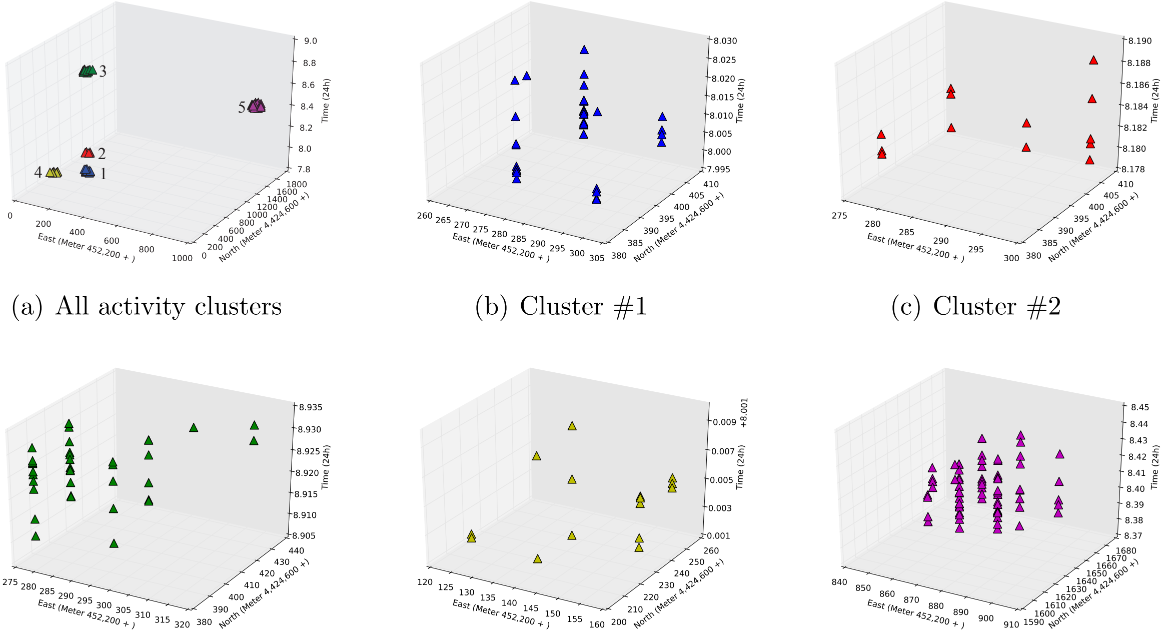
There are five activity clusters in Figure 11a. Each of them is zoomed in and showed in Figure 11b, 11c, 11d, 11e and 11f, where we can see how the clustered single activities look like in time and space. Three of these activity clusters (#1, 2 and 3) happened at the same location but in three different periods. We may infer that this user visited the same location three times per day. For example, some people leave home in the morning, return home at noon for lunch, and leave home again after lunch on weekdays; or she/he changed the activity (the time of visiting the place is changed). For instance, some bus drivers work in the night, followed by a day-shift later on. In this case, the location of bus pick-up is the same but the time is different. On the other hand, activity cluster #4 is likely similar to activity cluster #1 according to the method of detecting similar activity clusters proposed in this paper; thus, an activity change may have happened. Moreover, activity cluster #5 has high single activity density (Figure 11f), which means that user visited this place during certain time in a high probability.

With regard to user #128, as shown in Figure 12a, no activity clusters seems to be at the same location but in different times. However, activity clusters #1 and #4 are found to be similar. These two activity clusters do not have high single activity density (Figure 12b and e), which means this type of activity did not happen very often. On the other hand, the activity clusters with the highest single activity density, i.e., activity clusters #3 and #5, do not have similar activity clusters; thus, we could infer that there is not any significant activity change happened. The whole mobility pattern should be more stable than that of User #153.

Coincidently, the activity cluster distributions of these two users match the single activity distributions shown in Figure 9b, which has obvious activity change and in Figure 9c which always has the three identical activities in general.

4.3. Predictive results and evaluation

The activity clusters during January 2010 to June 2011, with and without implementing activity change detection (Algorithm 1), are used to train the predictive model. After implementing the predictive model, the activity clusters during July 2011 to September 2011 are used to evaluate the results. We calculate the transition probabilities between the activity clusters from the dataset during July 2011 to September 2011, which are treated as



(a) All activity clusters

(b) Cluster #1

(c) Cluster #2

(d) Cluster #3

(e) Cluster #4

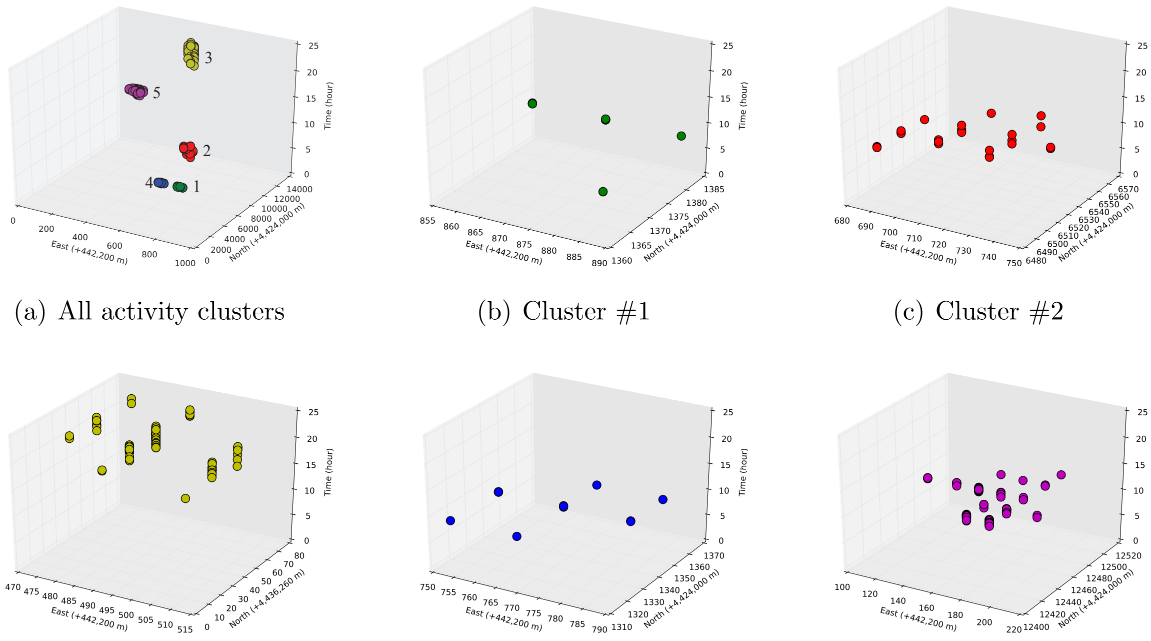
(f) Cluster #5

Figure 11: Activity clusters of user #153 (graphs scaled based on the clusters): (a) five cluster samples. (b) - (e) single activities in the cluster #1 - 5 shown in (a)

the evaluation benchmark. The results of the predictive model without considering activity change detection are used to compare with the results that consider activity change detection.

We select two activity clusters of A and B and list their predictive results and the real transition probabilities in Table 2 for user #153. Comparing the predictive results between no activity detection and Algorithm 1, the transition probabilities from activity cluster A to B, D and F are set to 0 with conducting Algorithm 1, which well matches the real transition probability during July 2011 to September 2011. In other words, the activity cluster B, D and F are changed activity clusters. Although these activity clusters happened during the early period, even some activity clusters are visited in a very high frequency from activity cluster A (e.g., the highest transition probability from A to B), the changed activity cluster are successfully removed by the Algorithm 1 that reduced the negative impact on predictive results. However, a new activity cluster (H) is involved in the visited activity clusters after July 2011, which cannot be detected. In fact, this type of case happens only if the new activity cluster comes later after model training. On the other hand, a similar activity cluster, F (new), of activity cluster F is successfully detected. Therefore, the transition probability of F (new) is 0.19 in Algorithm 1 instead of 0.12. Based on the discussion of clustering results, we believe that activity clusters F and F (new) correspond to activity clusters #1 and #4 in Figure 11 (a).

With respect to user #128, since the activity pattern of this user is stable overall, the results match the benchmark data very well even without performing change detection (Table 3). However, there still exist similar activity clusters with low single activity density (Figure 12c and Figure 12f). In other words, for those users whose activities are not changed in



(a) All activity clusters

(b) Cluster #1

(c) Cluster #2

(d) Cluster #3

(e) Cluster #4

(f) Cluster #5

Figure 12: Activity clusters of user #128: (a) five cluster samples. (b) - (e) single activities in the cluster #1 - 5 shown in (a)

general (refer to Figure 9c), they possibly still have activity change but only during a short period. Furthermore, the transition from B to D in Table 3 only happened during July 2011 to September 2011, which is not part of the training dataset. Therefore, the transition is not detected by the predictive model based on the dataset from January 2010 to June 2011, which leads the predictive model to predict this transition in a wrong way.

The R^2 values of these two predictive results based on the transition probability from July 2011 to September 2011 are computed for both users #153 and #128. For the predictive results without conducting activity change detection, the R^2 is 0.295 for user #153, while the R^2 is 0.965 for user #128. For the predictive results with consideration of the Algorithm 1, the R^2 is 0.762 for user #153, while the R^2 is 0.971 for user #128. Apparently, the R^2 has an obvious improvement from 0.295 to 0.762, when Algorithm 1 is used to handle activity changes of human mobility. For users that have stable activity pattern, the predictions with and without implementing change detection both have a very good performance.

5. Conclusion and future work

In this paper, we explore the spatio-temporal characteristics of human activities extracted from raw GPS trajectories using the Microsoft GeoLife dataset. Since the majority of human mobility predictions do not consider activity changes, we attempt to fill this gap by proposing a method to detect activity changes. A predictive model taking into account activity changes and built upon the concept of Markov chains is developed. This predictive model considers the spatio-temporal impacts on Markov chain states, which leads to a more

Table 2: Transition probability between activity clusters of user #153

Transitions	**Jan 2010 to Jun 2011**	**Jan 2010 to Jun 2011**	***Jul 2011 to Sept 2011***
	(no activity change detection)	**(Algorithm 1)**	
A → B	0.377	0.000	*0.000*
A → C	0.273	0.704	*0.492*
A → D	0.190	0.000	*0.000*
A → E	0.080	0.196	*0.253*
A → F	0.040	0.000	*0.000*
A → G	0.040	0.100	*0.000*
A → H	0.000	0.000	*0.255*
B → A	0.692	0.692	*0.604*
B → F	0.118	0.000	*0.000*
B → F(new)	0.070	0.188	*0.198*
B → G	0.120	0.120	*0.000*
B → H	0.000	0.000	*0.198*

Table 3: Transition probability between activity clusters of user #128

Transitions	**Jan 2010 to Jun 2011**	**Jan 2010 to Jun 2011**	***Jul 2011 to Sept 2011***
	(no activity change detection)	**(Algorithm 1)**	
A → B	0.843	0.843	*0.900*
A → C	0.157	0.157	*0.100*
B → A	0.155	0.155	*0.100*
B → C	0.745	0.745	*0.702*
B → D	0.000	0.000	*0.099*
B → E	0.050	0.000	*0.000*
B → E(new)	0.050	0.100	*0.099*

precise prediction. We implement the proposed predictive model for two users selected through visually analyzing the GeoLife dataset to evaluate its performance. The results show that the R^2 value is improved from 0.295 to 0.762 for the user with obvious activity changes, and 0.965 to 0.971 for the users without obvious activity changes. Thus, our proposed method, in terms of modeling activity changes, clustering activities and predicting human movement, presents an effort towards effective ways of improving the accuracy in analyzing and predicting human movements, and should provide new angles of studying the similar problems related to urban studies, e.g., urban planning, transportation modeling, and emergency response.

The study results are limited by the spatial and temporal coverage of the dataset used. With better quality data, the same method can be applied to predict human movement by different days of the week, the model results can be evaluated using the data from the same period of the year to avoid influence of different seasons, and some of semantic information may be inferred. The research work will further utilize the real-time human trajectories extracted from social media and investigate the event stream from sensor web platforms. On one hand, these real-time human trajectories can be used to keep on updating human mobility patterns, which is expected to provide a way to tackle the limitations caused by the high frequent activity changes of some people and new activities come right after the end of training samples. On the other hand, event streams are considered as a real-time input of the predictive model to sense what is happening around any specific users, and hopefully to predict certain related random events regionally and globally.

Acknowledgements

This work has been funded by the Natural Science and Engineering Research Council of Canada (NSERC). The authors would like to thank Professor Arnold Yuan for his valuable comments, Dr. Wai-Yeung Yan for his help in developing the manuscript, and anonymous reviewers for their constructive comments, which helped improve the paper.

References

Asahara, A., Maruyama, K., Sato, A., Seto, K., 2011. Pedestrian-movement prediction based on mixed markov-chain model. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, pp. 25–33.

Ashbrook, D., Starner, T., 2003. Using gps to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7 (5), 275–286.

Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. Nature 439 (7075), 462–465.

Chen, J., Shaw, S.-L., Yu, H., Lu, F., Chai, Y., Jia, Q., 2011. Exploratory data analysis of activity diary data: a space–time gis approach. Journal of Transport Geography 19 (3), 394–404.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Vol. 96. pp. 226–231.

Etter, V., Kafsi, M., Kazemi, E., Grossglauser, M., Thiran, P., 2013. Where to go from here? mobility prediction from instantaneous information. Pervasive and Mobile Computing 9 (6), 784–797.

Gambs, S., Killijian, M.-O., del Prado Cortez, M. N., 2012. Next place prediction using mobility markov chains. In: Proceedings of the First Workshop on Measurement, Privacy, and Mobility. ACM, p. 3.

Gonzalez, M. C., Hidalgo, C. A., Barabasi, A.-L., 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.

Jiang, B., Liu, X., Jia, T., 2013. Scaling of geographic space as a universal rule for map generalization. Annals of the Association of American Geographers 103 (4), 844–855.

Liao, L., 2006. Location-based activity recognition. Ph.D. thesis, University of Washington.

Markov, A., 1971. Extension of the limit theorems of probability theory to a sum of variables connected in a chain.

Mathew, W., Raposo, R., Martins, B., 2012. Predicting future locations with hidden markov models. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, pp. 911–918.

Song, C., Koren, T., Wang, P., Barabási, A.-L., 2010a. Modelling the scaling properties of human mobility. Nature Physics 6 (10), 818–823.

Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010b. Limits of predictability in human mobility. Science 327 (5968), 1018–1021.

Song, L., Kotz, D., Jain, R., He, X., 2006. Evaluating next-cell predictors with extensive wi-fi mobility data. IEEE Transactions on Mobile Computing 5 (12), 1633–1649.

Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.-Y., 2011. Recommending friends and locations based on individual location history. ACM Transactions on the Web (TWEB) 5 (1), 5.

Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y., 2009. Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th international conference on World Wide Web. ACM, pp. 791–800.