A Hierarchical Hidden Semi-Markov Model for Modeling Mobility Data

Mitra Baratchi, Nirvana Meratnia, Paul J.M. Havinga
Pervasive Systems Research group
University of Twente
Enschede, The Netherlands
{m.baratchi,n.meratnia,p.j.m.havinga}@utwente.nl

Andrew K. Skidmore, Bert A.K.G. Toxopeus
ITC
University of Twente
Enschede, The Netherlands
{a.k.skidmore,a.g.toxopeus}@utwente.nl

ABSTRACT
Ubiquity of portable location-aware devices and popularity of online location-based services, have recently given rise to the collection of datasets with high spatial and temporal resolution. The subject of analyzing such data has consequently gained popularity due to numerous opportunities enabled by understanding objects' (people and animals, among others) mobility patterns. In this paper, we propose a hidden semi-Markov-based model to understand the behavior of mobile entities. The hierarchical state structure in our model allows capturing spatio-temporal associations in the locational history both at stay-points and on the paths connecting them. We compare the accuracy of our model with a number of other spatio-temporal models using two real datasets. Furthermore, we perform sensitivity analysis on our model to evaluate its robustness in presence of common issues in mobility datasets such as existence of noise and missing values. Results of our experiments show superiority of the proposed scheme compared with the other models.

Author Keywords
Hidden semi-Markov model; mobility data analysis; movement modeling; movement prediction; next place prediction; Big data analytics

ACM Classification Keywords
H H.2.8 [Data Management]: Database Applications – Data Mining

INTRODUCTION
Thanks to the emergence of ubiquitous location-aware personal devices such as smart-phones, large volumes of mobility data are being collected every day. In spite of general worries regarding privacy concerns, surprisingly, people do not show much hesitation to (at least partially) share their location data through various online location-aware applications, services, and social networking websites [1] (for instance, through location tags in Facebook, Twitter, and Foursquare).

UbiComp'14, September 13–17, 2014, Seattle, WA, USA
Copyright 2014 ACM 978-1-4503-2968-2/14/09...$15.00.
http://dx.doi.org/10.1145/2632048.2636068

Mining such data has become an interesting research topic during the past decade due to its public and personal benefits in various domains. Potential application areas incorporating these benefits range from animal migration analysis, urban planning, and disaster relief [2] to mobile advertising, and even improving opportunistic routing algorithms [3]. Recently, researchers have used tourist mobility data for analyzing crowd dynamics during festivals to avoid security issues [4]. Management of traffic delays and congestions [5] is also another example application. From personal point of view, different kinds of location-based services and applications are becoming popular as they provide a better quality of private and professional life for people [6, 7]. Apps such as Moves [8] learn people's habits, and lifestyle through analyzing their location data. Athletes are also showing interest in location-aware applications such as Mytracks app [9] to improve their physical status by tracking their location combined with other physiological and activity measurements during exercises. In smaller scale location prediction can be used in identification of people in smart homes [10].

What all these applications and services have in common is their dependency on availability of knowledge about behavior of mobile entities. Having such knowledge helps predicting future mobility patterns, as well as identification of abnormal occurrences in the current patterns. A detailed movement model, which is more than extracting patterns of visit to a number of frequently visited places, can greatly contribute to acquiring such knowledge.

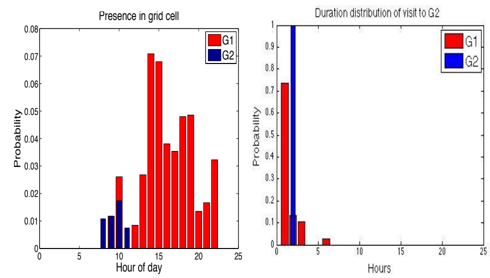
Various spatio-temporal rules and dependencies are hidden in mobility data caused by different types of context variables such as type and frequency of activities performed, and means of transport between stay-points (or regions of interest). A detailed model should encompass all these rules and dependencies. To better elaborate the spatio-temporal rules hidden in mobility data, an example from a real dataset is shown in Figure 1. Two visited grid cells, denoted by G_1 and G_2, have been extracted from a user's trajectory $\{o_t \mid t \in [1, T]\}$in Geolife dataset [11-13]. Figure 1.a represents the probability $\left(P_h^{G_i}\right)$ of user's presence (o_t) in these two different grid cells during different hours (h) of day ($P_h^{G_i} = \frac{|S_h^{G_i}|}{|S_h^+|}$ if $S_h^{G_i} = \{|o_t||o_t = G_i \& \ t \ mod \ 24 = h\}$ and $S_h^+ = \{|o_t||o_t = G_{j\in 1\ldots n} \ \& \ t \ mod \ 24 = h\}$). Figure.1.c is the probability that visit to one place is followed by visit to another place in one hour and Figure.1.b represents the probability distribution of the duration of visits to these cells.

A number of rules can be extracted from these images. It can, for instance, be seen from the duration distribution graph that presence of the user in each of these grid cells has a somewhat normal distribution (rule of duration of visits). Another interesting point is existence of a dependency in visits to these places. Although presence in each of these grid cells is of low probability, all visits to grid cell G_2 are followed by visit to grid cell G_1, while no visit to G_1 has been followed by a visit to G_2 (rule of transition between grid cells). These are important pieces of information, which can be used to efficiently model movement of this particular entity. This model can be used later for different purposes such as predicting future movements or to identify changing points in the mobility habits.

Designing a model, which can capture all the above-mentioned dependencies from real-life mobility datasets, is a challenging task. Firstly, trajectories are formed by components with different speeds (stay-points and transitions) being repeated with different frequencies. A model, which only captures frequency of visit to places, turns out to be biased to stay-points. On the other hand, preprocessing trajectories to take out segments with similar speed is time and energy consuming. Secondly, due to the inherent limitations of mobile data acquisition and collection methods, mobility data are extremely sparse and noisy. The sparseness is sometimes caused by the system designer as a tradeoff between accuracy and life-time requirements and sometimes it is caused by technical issues such as device mal-function. Mobility data is also noisy due to multipath and atmospheric effects.

In this paper we propose a model, to understand movement behavior of mobile objects. More explicitly, our contributions are as follows:

- We propose a hierarchal hidden semi-Markov-based model (HHSMM) which can capture both frequent and rare mobility patterns in the movement of mobile objects. Such model, in the first place, can be used in understanding mobility behavior of mobile entities such humans and animals. It can also be used in future movement prediction, which is an essential requirement in many ubiquitous applications such as urban planning, disaster relief, animal migration analysis, and mobile advertising.
- We apply the proposed model on two real datasets and show how the model can find such patterns (e.g. frequent, rare, weekly) without a-priori knowledge about mobile object's behavior.
- We evaluate the performance of our model in terms of its correctness in prediction of mobility behavior and compare it with other spatio-temporal models.
- We also, test the sensitivity of the proposed model in presence of noise and missing measurements, which are inherent characteristics of mobility datasets.



	G1	G2
G1	-	1
G2	0	-

(a) (b) (c)

Figure 1. An example depicting spatio-temporal rules in a mobile object's history, (a) presence probability in the grid cell over 24 hours, (c) duration distribution of visits, (b) transition probability from one place to another.

The rest of this paper is organized as follows. The second section presents the related work. Background information on hidden Markov models are presented in the third section, while our proposed hierarchical model is explained in the fourth section. The performance evaluation using both synthetic and real datasets are presented in the fifth section, and a number of remarks in the last section concludes the paper.

RELATED WORK

The problem of mining different mobility patterns from trajectories has been the subject of various research papers previously. For instance, location data have been analyzed to find frequent and popular routes [14], extract social context [15], trajectory clustering [16], or finding abnormal trajectories [17], to name but a few.

More recently, modeling movement has attracted attention in terms of models for movement prediction or frequent pattern detection [18-21]. Authors of [22], used semantic information to form semantically tagged trajectories in order to predict future mobility patterns. An ensemble method has been used by [23] to probabilistically model the movement on frequently visited places considering different context variables. Another group of researchers have used frequency domain analysis techniques to find the periodic patterns bound to activities of humans and animals. Both long-term and short-term periodic movements are found and presented. The periodic pattern mined in this way is used in prediction and abnormality detection [24, 25]. Authors of [26] have used the topic model Latent Drichlet Allocation to learn long duration sequences. Another approach proposed in [19, 27] is to use principal component analysis and Eigen value decomposition to model the movement pattern. Typical pattern of different types of days (Weekdays, weekends, and public holidays) can be discovered by this technique.

Irregular but highly ordered events that happen less often, yet with strong dependencies, are not effectively considered in the above models. This makes them unsuitable for cases, where online prediction of mobility behavior in near future is crucial. Having a very specific attribute of movement (such as semantic information, periodicity, Eigen values) in all the methods mentioned above, make them very specific to certain frequent patterns.

Recently, a number of state-space models have been proposed to model movement of mobile objects. These models attempt to capture the variation in spatial dependencies.

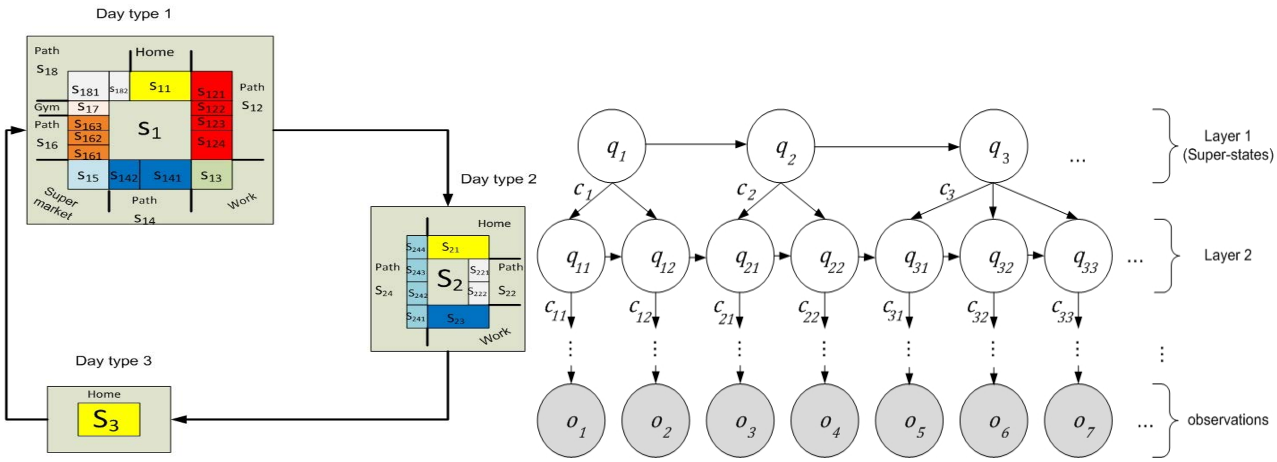
Markov model is one of the well-known state-space models. Different version of this model has so far been applied on mobility data. For instance, Order-k Markov model was used in [28] to predict movement of users in Wi-Fi network cells. In [20], a model is proposed based on hidden Markov models for modeling movements from one stay-point to another, while authors of [29] used mixed Markov model for the same purpose. A mixed autoregressive hidden Markov model is proposed in [30] on stay-points. The main drawback of these methods is that they do not completely consider the temporal variability in the mobility data. In these models, a trajectory is only partially used either as a sequence of visited stay-points or just as the transition path between stay-points. Apart from being incomplete, these methods require pre-processing the data to extract regions of interest or stay-points. The required pre-processing phase is rather time/energy inefficient. The above-mentioned problem is due to the inherent limitation of hidden Markov Model which considers constant duration for each system state. Therefore, there is still need for a model, which can be applied on complete mobility data, consisting of both stay-points, and transitions by considering their temporal variability. Hidden semi-Markov model addresses the above-mentioned issue by considering an additional duration property for each state. To the best of our knowledge there is only one previous research [31] which has considered using hidden semi-Markov model on mobility data. However, the authors have only evaluated their model on a synthetic dataset representing data of few hours. As we show when modeling large dataset of human mobility, composed of complex patterns (e.g. weekly), the technique used in [31] results in a very course grained model. To address this issue we propose a hierarchical hidden semi-Markov model offering the following advantages: (i) the fine-grained structure of the model provides higher accuracy; (ii) its granularity is adjustable to resources available, and (iii) hierarchical structure of the model improves the speed of parameter estimation.

BACKGROUND

State space models

We begin this section by providing background information on different state space models and their parameters. In the original hidden Markov model, due to the first order Markov assumption, it is implicitly assumed that the duration of each system state is constant or exponentially distributed. As a consequence, in these models transition between states happens at any time and self-transition is allowed. The downside of this model is that it does not take any advantage of the information hidden in the duration probability of visit to different places. Stay-points and transition paths have different duration distributions, which also need to be taken into account. In order to deal with this problem, later the original hidden Markov model was extended to hidden semi-Markov model, where apart from the transition between states there is an additional parameter for explicitly modeling the duration of states. The hidden semi-Markov model (also known as explicit duration hidden Markov model or variable duration hidden Markov model) is represented by $\lambda = (Q, O, A, B, C, \pi)$. In the above model, $O = \{o_t | t \in T\}$ and $Q = \{q_t | t \in T\}$ represent the entire observation sequence, and the entire high level state sequence, respectively (T is the set of uniformly distanced timestamps). A is the $M \times M$ state transition probability matrix representing the probability of change between states expressed as ($a_{ij} = p[q_{t+1} = s_j | q_t = s_i]$). B is $M \times N$ emission probability matrix representing the conditional probability between states and observations ($b_i(v_j) = p[o_t = v_j | q_t = s_i]$), and π is the initial probability distribution vector of size $M \times 1$, ($\pi = p[q_1 = s_i]$). C is the additional important *($M \times D$)* matrix added to the previously mentioned parameters of hidden Markov model where D is the maximum state duration and ($c_i(d) = p[c_{s_i} = d]$) represents the probability of state s_i last for d time units. This type of model is previously used for presenting a sequence of events with different durations; for instance, in video image processing, and daily activity modeling [32].

Given an output sequence in form of a sequence of observations, a parameter-learning algorithm is performed to estimate the parameters of the model λ. The best set of state transition, output probabilities, and state duration matrices is estimated in this way. One of the well-known decoding algorithms used for this purpose is Baum-Welch algorithm [33]. Among different variations of Baum-Welch used for modeling with hidden semi-Markov model, we have chosen the method proposed in [31], as it considers missing observations.



(a) (b)

Figure 2. (a) Hierarchical structure in mobility data, (b) Graphical model representing hierarchical hidden semi-Markov model

PROBLEM OF HIDDEN SEMI-MARKOM MODEL (HSMM) AND OUR HIERARCHICAL SOLUTION (HHSMM)

In this section, we explain our proposed hierarchical model for modeling movement. Let us assume that the movement dataset $O=\{o_{t_i} \dots o_{t_n}\}$ of a moving object over several days is given. This dataset is composed of chronologically ordered two-dimensional geo-spatial points representing object's location at time-stamp i, where the distance between t_i and t_{i+1} is variable. We are looking for a model λ composed of a number of spatio-temporal rules relating the noisy and unevenly sampled movement data to their context related state Q=$\{q_{t_i} \dots q_{t_n}\}$. These states can represent the activity governing the movement.

Our aim is to model the complete movement track in a way that each state in the model is either a stay-point or the transition path from one stay-point to another, where spatial coordinates have some form of spatio-temporal similarity. We assume that the sequence of places that the person visits is a Markovian process with hidden states being the context ruling person's activities and the places that a person visits being observable two-dimensional spatial points. A possible solution would be to consider each spatial point as an observation and use hidden semi-Markov model to find the most probable sequence of states that explain the observations.

However, when a series of behaviors are repeated periodically (for example, over a day) they will be found as one super-state, while the whole super-state might be composed of smaller states. This way the desirable granularity, which is required to relate observations to concepts such as stay-points and paths, is not provided. In fact, the spatial points, which are closer to each other spatially (i.e. stay-points) or spatio-temporally (i.e. paths) are more probable to belong to one state. Therefore, as the state duration distribution of different activities are different, simply considering each observation as a spatial point is not enough to find the states that are explainable with human logic. The problem is better explained in Figure 2.a. In this figure, the repetitive behavior of a person is illustrated. This behavior is consisted of three super-states with duration of a day. As it can be seen, each super-state is also composed of a number of smaller states, which represent visit to different places each with different durations. As will be shown in section 5, by using HSMM only the higher level states are found giving a very high granular view of the movement pattern where sometimes a complete day is discovered as a single state.

We found the following problem with the original HSMM:

The original HSMM treats observations as nominal values. Thereby, there is no consideration for the distance between the observations.

In order to solve the problem mentioned above, we propose using a hierarchical hidden semi-Markov model taking into account the distance between observations in each state. This hierarchical model is defined as $\lambda = (Q, O, A, B, C, \pi)$ such that each state s_i in the model is itself a hierarchical hidden semi-Markov model $\lambda_{s_i}^h$, $(h > 1)$:

$$\lambda_{s_i}^h = (Q_{s_i}^h, O_{s_i}^h, A_{s_i}^h, B_{s_i}^h, C_{s_i}^h, \pi_{s_i}^h)$$

The new observation sequence $O_{s_i}^h$ is only composed of the observations which were categorized into one higher-level state $O_{s_i}^h$=$\{o_t | q_t^{h-1} = s_i^h\}$. Each state in the final level (h) is only composed of observations, which are spatially close to each other. Figure 2.b visualizes our proposed hierarchical hidden semi-Markov model through a graphical model.

Algorithmic details

Algorithm 1 summarizes the procedure of training our hierarchical model.

Algorithm 1 HHSMM

INPUT: Maximum number of states in each level $M_{1..h}$, distance threshold th, Maximum state duration $D_{1..h}$, O observation sequence

OUTPUT: State transition probability matrices $A^h_{s_i}$, Emission probability matrices $B^h_{s_i}$, State duration probability matrices $C^h_{s_i}$, initial probability matrices $\pi^h_{s_i}$

ALGORITHM:

1 $[A, B, C, \pi, Q]$=TrainHSMM(M_1, D_1, O); // Train the basic level HSMM;

2 **For** $i = 2$ to h **do**

3 $CandidateStates$= all states found in previous level;

4 **While** $CandidateStates$ is not empty **repeat**

5 Remove any state from $CandidateStates$ with points lying within a

6 circle with radius th;

7 **For** j = 1 to length($CandidatesStates$) **do**

8 $O^i_{s_j}$={$o_t | q^{i-1}_t = s^{i-1}_j$ };

9 $[A^i_{s_j}, B^i_{s_j}, C^i_{s_j}, \pi^i_{s_j}, Q^i]$=TrainHSMM($M_i, D_i, O^i_{s_j}$);

The algorithm gets the input sequence and models the mobility pattern. Due to various environmental (such as cloud cover) and technical reasons (such as device malfunction), it is improbable that equal coordinates are reported for one place. Therefore, we firstly map location coordinates into cells of a gridded map where each observation is replaced by the relevant cell id, where it is located. The algorithm further proceeds as follows. First, hidden semi-Markov model is used to model the input sequence and to find the super-states in the model (line 1). It is probable that regular days with similar repetitive sequence of places being visited are found as one state. To have a higher resolution insight, in case in each of these high level super-states, there are observations with a distance greater than a threshold then that state will be chosen for being remodeled. On the next step, we apply hidden semi-Markov model on each of these states (lines 7-9). This step can be repeated until no other states with such condition are found. It should be mentioned that, the algorithm for training HSMM (TrainHSMM) is not presented due to space limitation but the interested user can refer to [34].

Complexity analysis

Complexity of the light Baum-Welch training algorithm [34] TrainHSMM is $O(MD + M^2)T$ and the memory required for its training is $O(MT)$. Like all hidden Markov based algorithms, in case a large number is chosen for the states and their duration (the maximum "naïve" number for M and D is the number of unique observations, and length of observation sequence, respectively), the algorithm becomes computationally expensive. This, however, is not the case for our HHSMM algorithm. As shown in [35], there is high degree of temporal and spatial regularity in human trajectories, and each individual can be characterized by a significant probability of returning to a few frequently visited locations. Due to this, a high degree of people's acts can be summarized using very little number of super-states, which can be analyzed in more detail in case of necessity. The advantage of this hierarchical model is that its complexity is adjustable. It is not required that the number of states are initially set equal to all unique observations. A limited number of states, with longer durations for the higher levels can be used. In each iteration of the algorithm the number of states M_h increases while the parameters D_h and T_h decrease, leaving the complexity of learning for each intermediate state in each level balanced $O(M_h D_h + M_h^2)T_h$, $(T_h < T_{h-1}, D_h < D_{h-1}, M_h > M_{h-1})$. Therefore, the model can be efficiently trained with respect to the resources available and the granularity required. The hierarchical model, also gives the possibility of further improvements in sampling frequency and resolution of observations for each level. In higher levels the number of super-states is limited and low frequency sampling is enough. By adjusting the size of the grid based on the movement area, the number of distinct observations will be reduced requiring less number of states for higher-level states.

PERFOMANCE EVALUATION AND COMPARISON

Datasets

For our experiments we have chosen the following real datasets:

Geolife dataset [11-13]*:* This dataset is collected in GeoLife project organized by Microsoft Research. GPS trajectories in this dataset, collected by 165 users, have various sampling rates. For the majority of users the sampling frequency is as high as reporting a sample every five seconds. However, there are considerable long intervals (hours) when no data is available due to various reasons such as device mal-function or intentional turning off of the device. The users have recorded mobility data during various activities and habits of their daily life in this dataset.

Capricorn dataset [36, 37]*:* This dataset is composed of GPS data collected by data loggers attached to a Capricorn. The Cretan Capricorn (Capra aegagrus-cretica) lives in the White Mountains and is endemic for Crete. Due to increasing livestock populations (goats) the population is threatened. As the species is difficult to locate very little is known about their habitat use in different seasons. Since mid-July of 2011 one male and two female Capricorns have been equipped with GPS collars. By deploying animal collars equipped with GPS, precise spatio-temporal data are provided in small time intervals [38]. The daily 16 GPS-fixes acquisition schedule had very short intervals in the morning (08:00 – 11:00) and in the afternoon (20:00 – 23:00) based on the daily behavior of the Capricorn and because they show more activity at these parts of the day [39].

Three moving entities (two people, and one capricorn) have been chosen from these datasets with three different movement profiles. Table 1 summarizes the parameters of the movement profile of each of these moving entities. As seen, these three cases represent three general movement profiles which are 1) high range movement, high average speed, 2) medium range movement, medium average speed, and 3) low range movement, low average speed. We see that for the second user, both maximum speed and movement area are large numbers. The maximum speed is in range of airplane speed (254 Km/h), which can also be explained by a number of coordinates in the dataset, which are in proximity of the airport.

Parameter	Moving Object		
	Geolife User 1	Geolife User 2	Capricorn
Movement area (km^2)	76.6	5.2×10^5	2.8×10^3
Total disp	1.4×10^3	9.7×10^3	2.6×10^4
Dt (days)	76	254	332
Average speed (km/h)	5×10^{-5}	0.24	0.08
Max speed	71.6	240	2
Missing	76%	88%	71%

Table 1. Movement profile of the moving entities.

Case studies on Geolife dataset

In order to show the process of training the hierarchical hidden semi-Markov model we show the procedure of building a two-level hierarchical model with algorithm 1 on first and second user of Geolife dataset.

User 1

Figure 5.a shows that the mobility data of User 1 is more concentrated in a small area. In order to train the HHSMM, we chose the values 168, and 24 for maximum state duration of first and second level, 10 for the number of states in each level, with a grid size of 10×10. After being gridded, all the points (observations) falling in the same cell are assigned with a unique number (Figure 3.b). In the first level, two super-states are found with different duration distributions shown in Figure 3.c (while we chose number 10 for the number of states, after training the rest of the states were not assigned to any observation). These two duration distributions, with means near 120 and 48 hours, evidently represent the general distinction in mobility behavior of this person in weekdays and weekends. This is an interesting positive characteristic of our model, as it can find proper duration distribution without us making any assumption on this typical weekly behavior. Such patterns were previously found with complex periodicity analysis [24, 25]. In order to see if this separation is valid we have also presented the probability of visit to different grid cells (G_i) over hours (h) of week in Figure 3.a ($P_h^{G_i} = \frac{|S_h^{G_i}|}{|S_h^{+}|}$ if $S_h^{G_i} = \{|o_t| \,|\, o_t = G_i \,\&\, t \bmod 168 = h\}$ and $S_h^{+} = \{|o_t| \,|\, o_t = G_{j\in 0\ldots n} \,\&\, t \bmod 168 = h\}$) ($G_i = 0$ represents missing data). As seen, some of the cells are visited "mainly" in the first 120 hours (2,3,12,13), some have only been visited in the last 48 hours (5,14,15) and some in both (1,4,6,7,11).

After remodeling is done, super-states 1 (weekends) and 2 (weekdays) are modeled by 4 and 2 numbers of lower-level states, respectively. Points that are assigned to each state with high probability are represented in Figure 3.d, 3.f ($p[q_t = s_j | o_t = v_i]$) with "$*$" sign. Those points which are drawn in a circle represent the observation, which represents states with highest probability ($p[o_t = v_i | q_t = s_j]$).

Looking at the duration distributions, and probability of presences (Figure, 3.a, 3.b, 3.e., 3.g), one can note that some of the states found are still dividable to more lower-level states. For instance, in state 3 of super-state 1 (shown in black), observations 1 and 15, are relatively far from each other and can still be divided into two separate states. Unfortunately, due to space limit we cannot represent all possible lower-level states. It should also be mentioned that hidden semi-Markov models, not having a rough guess about the emission and transition matrices, do not always converge to the same results. In order to have understandable states, we repeat the learning process few times to get state durations, which follow a normal distribution.

User 2

As it can be seen from figure 5.b, the mobility data of user 2 is composed of very long travel sequences. Using algorithm 1 with the previous parameters used for user 1 and grid size 100×100, we found two general super-states for this user as shown in Figure 4.b. The super-state colored in blue represents the points corresponding to long traveling sequences. Due to its rare nature (9% of dataset) and average high speed of the user in this state, most of the points in this state are only observed once (the median and mean of number of times each observation is observed in the dataset is 1 and 3, respectively). However, as it can be seen in Figure 4.c, after the observations in this super-state are remodeled, 4 states are found which can represent the ways to and from two stay-points, as well as the stay-points in different cities. With previous Markov based methods such visible states are impossible to find when only stay-points are used for modeling. The red super-state is a dense representation of points in an area where more than 91% of the observations are located. After the super-state is remodeled (Figure 4.d), two lower-level states are found. Although most of the points in this state are only observed once, the similarity between points in this state is that they are followed by visit to the points in the other state.

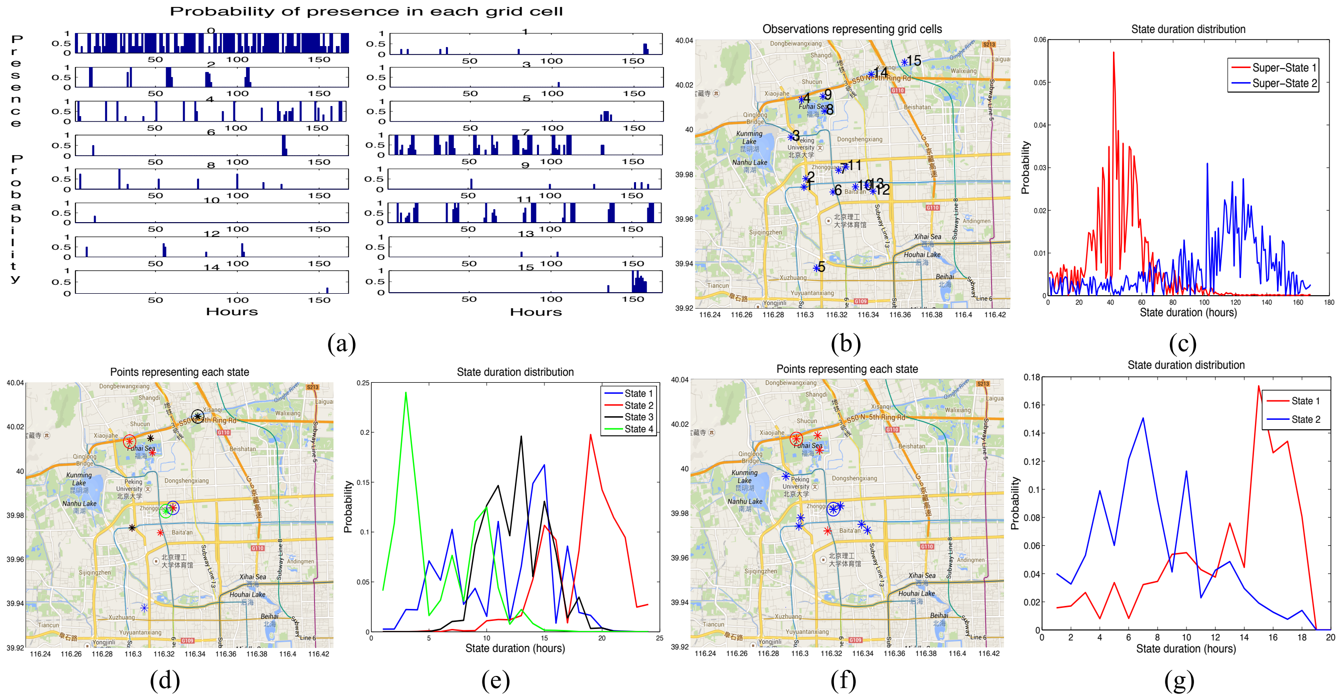


Figure 3. Modeling the mobility data of user 1 in Geolife dataset (a) Probability of presence in each grid cell, (b) Enumerated grid cells visited by user 1, (c) duration distribution of superstates, (d,e) modeling super-state 1, (f,g) modeling super state 2.

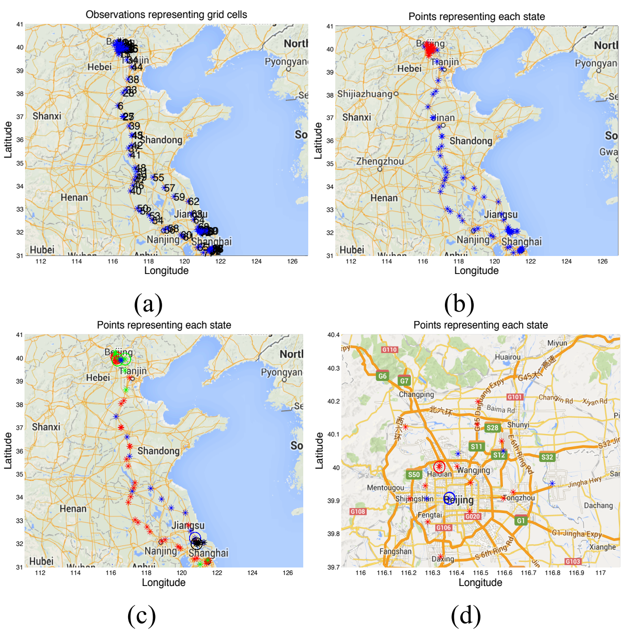


Figure 4. Modeling the mobility data of user 2 in Geolife dataset. (a) Enumerated grid cells visited by user 2 (b) Superstates (c,d) Second level states in each superstate.

Comparison with other models

In this section we compare our proposed model with the other models in literature, using both synthetic and real datasets. As most of the previous research using HMM require pre-processing, we have chosen HSMM and a number of spatio/spatio-temporal models, which can be directly applied to complete trajectories. In each of these models different spatial and temporal priors are used. Inspired by [18], the models used for performance evaluations are:

Spatial prior model (SP): In this model presence in each location depends on a prior location. SP is purely spatial and does not use any temporal context.

$$p^{SP}(o_t = v_i |\, t = t_i, o_{t-1} = v_j) = p(o_t = v_i | o_{t-1} = v_j)$$

Hourly prior model (HP): In this model presence in each location depends on its hourly visit distribution.

$$p^{HP}(o_t = v_i |\, t = t_i, o_{t-1} = v_j)$$

$$= p(o_t = v_i | t_i \bmod 24 = h)$$

Spatial-hourly prior model (SHP): In this model presence in each location depends on the hourly distribution, as well as the prior location:

$$p^{SHP}(o_t = v_i |\, t = t_i, o_{t-1} = v_j)$$

$$= p(o_t = v_i | o_{t-1} = v_j \,\&\, t_i \bmod 24 = h)$$

The hidden semi-Markov model (HSMM): This model is the basic hidden semi-Markov model where presence in each location depends on the current state, and the residual time of the states:

$$p^{HSMM}(o_t = v_i |\, t = t_i)$$

$$= p(o_t = v_i |\, (q_t, \tau_t) = (s_m, d))$$

The hierarchical hidden semi-Markov model (HHSMM): This model is the one proposed in this paper where presence in each location depends on a hierarchy of current states, and their remaining times:

$$p^{HHSMM}(o_t = v_i \mid t = t_i)$$
$$= p(o_t = v_i \mid \forall h,\ (q_t^h, \tau_t^h) = (s_m^h, d^h))$$

Evaluation in terms of correct prediction

In order to evaluate the models, we chose to test how we can use them to accurately predict near future events. Our analysis is composed of two phases:

- *Training phase*: First, all three datasets are equally sampled per hour forming a time-series where missing values are replaced by 0. Next, we divide each dataset into two parts. During training, the first half is completely given as input to algorithm 1. The maximum state duration is 168 and 24 hours, which represent states of maximum size of a week and a day. While these values are chosen with respect to the length of datasets used for training, longer durations for super-states can be used when the datasets are larger to find longer patterns. The number of states we chose for each level is set to 10. During tests, we observed that the number of states chosen is more than enough for all datasets, as some states are not assigned to any observation. The distance threshold used for algorithm to re-model a state is 1000 meters. After the model is trained, for HSMM and HHSMM models we calculate a $N \times M$ size matrix R which represents the relation between observations and states ($r_i(s_j) = p[q_t = s_j \mid o_t = v_i]$). This matrix is used in prediction.
- *Prediction phase:* we check predictability of the models on the second half of the dataset. For each two consecutive timestamps where data is not missing $\{\forall\ (i, i+1) \mid (o_i, o_{i+1})\ are\ not\ missing\}$, and o_{i+1} had been observed in the training dataset, we check to see how we can predict the data of the second timestamp (o_{i+1}) from the prior one (o_i).

The procedure was repeated 50 times with different grid sizes (varying from 10×10-500×500 for the first user, 500×500-1000×1000 for the second, and 1×1-50×50 for the capricorn). These sizes have been chosen based on the movement ranges.

Graphs shown in Figure 5 compare the efficiency of each of these models in terms of their prediction accuracy. The first two columns are the results of performing experiments on the human dataset and the last column is that of Capricorn data. In each column the movement range of the moving object after being sampled, total prediction accuracy, prediction of change accuracy, and cost of wrong prediction are shown. These parameters are explained bellow:

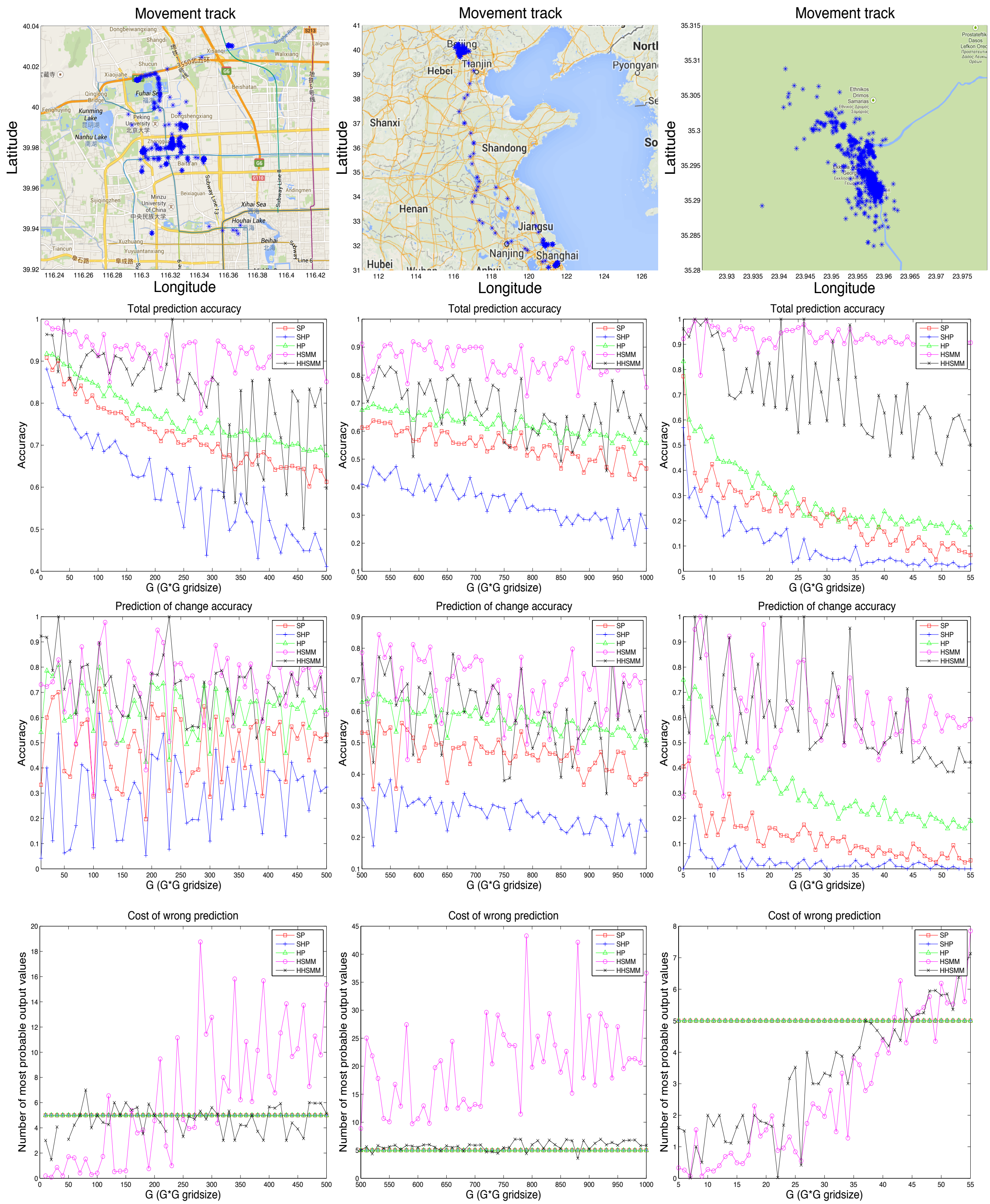
Total prediction accuracy: This graph represents the total correct predictions both when the next destination is in the same cell and when it is in another cell.

Prediction of change accuracy: As the periods of stay and movement are unequal, it is almost always easier to predict points where the moving object is stable (predicting the current spatial point as the next destination ($o(t+1) = o(t)$)). Therefore, as well as showing the accuracy of models in terms of total prediction, we also show the results of predicting the points which represent a change from the previous timestamp ($o(t+1) \neq o(t)$). This helps in showing the difference of the algorithms in predicting these two different types of measurements.

Cost of wrong prediction: This graph represents the number of cells proposed with highest probability for each wrong prediction. This will show the cost of each wrong prediction. The reason for showing this graph is that, for HSMM and HHSMM, it is possible that each state is composed of a group of observations. Therefore, by using the observation/state matrix (R) this group of points, belonging to one state, will be suggested as next point prediction having the same probability ranges. In the other models, however, the most probable point has a higher probability which can be used in prediction. In order to be fair, we also compare the methods in terms of the cost of this inaccuracy. As the cost of HHSMM is lower than HSMM, we adjusted the cost of the other models with this model by accepting more predictions. In this way for the other models we always accept the top 5 most probable points for predicting the next destination.

Looking at Figure 5, one notices the following remarks:

For the first two datasets, the HSMM model performs considerably better than all the other models in terms of total and prediction of change accuracy. This comes, however, with a considerable high cost for each wrong prediction. This shows the high granularity of the states that are the outcome of HSMM. HHSMM follows HSMM in total and prediction of change accuracy with a cost of wrong prediction being much lower than that of HSMM and in range of other methods. Prediction of change accuracy with these two methods is higher than the other methods. This is resulted by correct duration estimation for each state. For the Capricorn dataset, HSMM and HHSMM are very close in prediction of change accuracy and total prediction accuracy of HSMM is higher than HHSMM. However, this time the cost of wrong prediction of two methods is very close. This can explain that the animal's movement is less structured, and that the hierarchical structure has not been able to add to the accuracy. In this case, the higher granular model is successful.



(a) (b) (c)

Figure 5. Comparison of different models on 3 moving objects. Column a, b, and c represent graphs of user 1, user 2 and Capricorn, respectively.

In most cases by increasing the grid size accuracy decreases. This is due to an increase in the number of unique observations. The uneven shape of lines is due to discretizing observations to grid cells and is caused by the fact that the data had not been preprocessed. The other reason is that for HSMM and HHSMM the result of training is not always a unique model. Therefore, the best model is chosen after ten times of training.

Evaluation in terms of robustness against noise and missing values

In this section, we show the validity of our approach in presence of noise and missing values. For this purpose we use a synthetic dataset. This test helps checking the sensitivity of the above-mentioned models to these issues in a controlled setting. In order to produce the synthetic data, a movement generator was written with the parameters mentioned in Table 2.

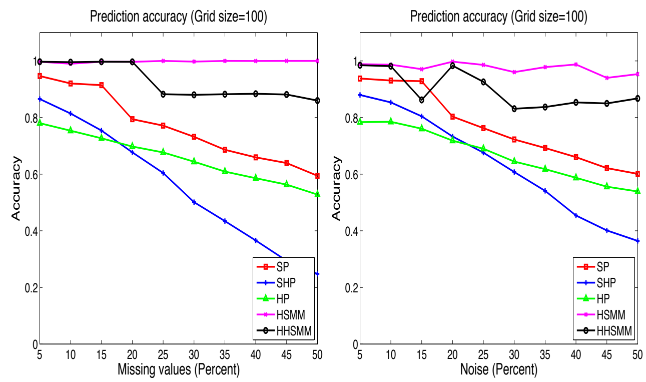
Parameter	Value
σ_{start}	120 min
σ_{end}	120 min
r	0.001
L	10
K	8
Missing samples (θ)	5-50% (N×24)
Noise (ρ)	5-50% (N×24)
Number of Grid cells	100×100
Total number of paths and places	7

Table 2. Parameters chosen for the test with synthetic dataset

This test sequence is composed of repetition of a sequence of geo-spatial points, which can represent a repetitive behavior of a person in visiting a number of places ($test_i = \{x_i, y_i | i \in [1, L\times 24]\}$). K number of places and the paths connecting them are chosen. The event of start and end of a visit to each of these places is expected to be at t_{start} and t_{end} and the actual visit happens within $t_{start} \pm \sigma_{start}$ and $t_{end} \pm \sigma_{end}$. After forming this sequence we test the effect of missing samples and noise with the tests mentioned bellow:

Test 1 (Missing samples): we generate θ random indexes and replace the indexing values $(x_\theta,\ y_\theta)$ by $(0,0)$ (representing missing observations). Next, we train each of the models on the resulting sequence. The success of each model is in correctly finding observations, which can replace each missing value.

Test 2 (Noise): we generate ρ random indexes and replace $(x_\rho,\ y_\rho)$ with a noisy value $(x_\rho + e_x,\ y_\rho + e_y)$ where e_x and e_y are randomly chosen from $[1, r]$. The parameters of the test are presented in Table 2. The success of each model is on correctly replacing the noisy observation with the original value.



(a) **(b)**

Figure 6. Success rate of algorithms in predicting the a) missing value and b) noise.

For the models SP, SHP, and HP, we chose the values, which had a probability over 0.5 for predicting the missing and noisy values. For HSMM and HHSMM we chose the cells which belonged to the state detected with probability more than 0.5.

As it can be seen from the Figure 6.a and 6.b, HSMM followed by HHSMM are superior to the other models both when missing values and noise are present in the dataset. Even when noise or missing values reach up to 50%, these two models perform considerably well. This is thanks to considering both forward and backward variables, which are able to find the best model representing the entire dataset. The accuracy of HSMM is higher than HHSMM as it predicts all points belonging to the super-states while HHSMM gives a finer grained predictions, albeit with a slightly reduced accuracy

CONCLUSION

In this paper the important and challenging problem of modeling movement tracks of mobile objects is addressed. A hierarchal hidden semi-Markov model is proposed to model the movement of mobile objects. This technique can model the movement both in stay-points and in the paths connecting them. As the evaluation results show, when applied to raw trajectories this method is robust in presence of noise and missing measurements. Furthermore, compared to the other models, this model can better capture dependencies in data to predict future movement patterns.

In our future work, we plan to extend HHSMM model with a streaming training mechanism so that it better meets limited resources of ubiquitous devices.

REFERENCES

1. S. Aflaki, N. Meratnia, M. Baratchi, and P. J. M. Havinga, "Evaluation of Incentives for Body Area Network-based HealthCare Systems," presented at the IEEE ISSNIP, Melbourne, Australia, 2013.
2. X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 Haiti earthquake," Proceedings of the National Academy of Sciences, vol. 109, pp. 11576-11581, July 17, 2012 2012.
3. O. Türkes, J. Scholten, and P. J. M. Havinga, "RoRo-LT: social routing with next-place prediction from self-assessment of spatiotemporal routines," in 10th IEEE International Conference on Ubiquitous Intelligence and Computing, UIC 2013, Vietri sul Mare, Italy, 2013.
4. U. Blanke, T. Gerhard, F. Tobias, and P. Lukowicz, "Capturing crowd dynamics at large scale events using participatory GPS localization," in 2014 Ieee Ninth conference on intelifent sensors, sensor networks and information processing (Issnip), Singapore, 2014.
5. A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "VTrack: accurate, energy-aware road traffic delay estimation using mobile phones," presented at the Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, Berkeley, California, 2009.
6. J. Bakker, L. Holenderski, R. Kocielnik, M. Pechenizkiy, and N. Sidorova, "Stess@Work: from measuring stress to its understanding, prediction and handling with personalized coaching," presented at the Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, Miami, Florida, USA, 2012.
7. M. Shoaib, H. Scholten, and P. J. M. Havinga, "Towards Physical Activity Recognition Using Smartphone Sensors," in Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), 2013, pp. 80-87.
8. (2013). Available: http://www.moves-app.com
9. (23 May 2013). My Tracks. Available: http://www.google.com/mobile/mytracks/
10. A. S. Crandall and D. J. Cook, "Using a Hidden Markov Model for Resident Identification," presented at the Proceedings of the 2010 Sixth International Conference on Intelligent Environments, 2010.
11. Yu Zheng, Xing Xie, and W.-Y. Ma, "GeoLife: A Collaborative Social Networking Service among User, location and trajectory," IEEE Data(base) Engineering Bulletin, 2010.
12. Y. Yingxiang, "Understanding human mobility patterns from digital traces," Master of science, Civil and environmental engineering, Massachusettes institute of technology, 2011.
13. Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," presented at the Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009.
14. M. Baratchi, N. Meratnia, and P. J. M. Havinga, "Finding frequently visited paths: dealing with the uncertainty of spatio-temporal mobility data," in IEEE ISSNIP, Melbourne, Australia, 2013.
15. M. Baratchi, N. Meratnia, and P. J. M. Havinga, "On the Use of Mobility Data for Discovery and Description of Social Ties," in Accepted to appear in "The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining", 2013.
16. J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," presented at the Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China, 2007.
17. L. Jae-Gil, H. Jiawei, and L. Xiaolei, "Trajectory Outlier Detection: A Partition-and-Detect Framework," in Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, 2008, pp. 140-149.
18. H. Gao, J. Tang, and H. Liu, "Mobile Location Prediction in Spatio-Temporal Context," in Mobile Data Challenge 2012 Newcastle, UK, 2012.
19. A. Sadilek and J. Krumm, "Far Out: predicting long-term human mobility," in Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 814-820.
20. W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden Markov models," presented at the Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, Pennsylvania, 2012.
21. Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on GPS data," presented at the Proceedings of the 10th international conference on Ubiquitous computing, Seoul, Korea, 2008.
22. J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," presented at the Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, Illinois, 2011.

23. T. M. T. Do and D. Gatica-Perez, "Contextual conditional models for smartphone-based human mobility prediction," presented at the Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, Pennsylvania, 2012.
24. Z. Li, B. Ding, J. Han, R. Kays, and P. Nye, "Mining periodic behaviors for moving objects," presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, 2010.
25. M. Baratchi., Nirvana Meratnia, and P. J. M. Havinga, "Recognition of periodic behavioral patterns from streaming mobility data.," presented at the Proceedings of 10th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Mobiquitous 2013, Tokyo, Japan, 2013.
26. K. Farrahi and D. Gatica-Perez, "Extracting Mobile Behavioral Patterns with the Distant N-Gram Topic Model," in Wearable Computers (ISWC), 2012 16th International Symposium on, 2012, pp. 1-8.
27. N. Eagle and A. Pentland, "Eigenbehaviors: identifying structure in routine," Behavioral Ecology and Sociobiology, vol. 63, pp. 1689-1689, 2009/09/01 2009.
28. L. Song, D. Kotz, R. Jain, and X. He, "Evaluating location predictors with extensive Wi-Fi mobility data," in INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies, 2004, pp. 1414-1424 vol.2.
29. A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement prediction based on mixed Markov-chain model," presented at the Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, Illinois, 2011.
30. A. Asahara, K. Maruyama, and R. Shibasaki, "A mixed autoregressive hidden-markov-chain model applied to people's movements," presented at the Proceedings of the 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, California, 2012.
31. S.-Z. Yu and H. Kobayashi, "A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking," Signal Processing, vol. 83, pp. 235-250, 2003.
32. T. L. M. v. Kasteren, G. Englebienne, et al. "Activity recognition using semi-Markov models on real world smart home datasets," J. Ambient Intell. Smart Environ., vol. 2, pp. 311-325, 2010.
33. L. R. Welch, "Hidden Markov Models and the Baum-Welch Algorithm," IEEE Information Theory Society Newsletter, vol. 53.
34. Y. Shun-zheng and H. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration hidden Markov model," Signal Processing Letters, IEEE, vol. 10, pp. 11-14, 2003.
35. M. C. Gonzalez, C. A. Hidalgo, and A.-L. AU - Barabasi, "Understanding individual human mobility patterns," Nature 2008.
36. M. Baratchi, N. Meratnia, P. Havinga, A. Skidmore, and B. Toxopeus, "Sensing Solutions for Collecting Spatio-Temporal Data for Wildlife Monitoring Applications: A Review," Sensors, vol. 13, pp. 6054-6088, 2013.
37. C. Katsaounis, "The Cretan Capricorn (Capra aegagrus-cretica) lives in the White Mountains and is endemic for Crete," Master of Science, ITC, University of Twente, 2012.
38. M. Hebblewhite and D. T. Haydon, "Distinguishing technology from biology: a critical review of the use of GPS telemetry data in ecology," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 365, pp. 2303-2312, July 27, 2010 2010.
39. M. C. Nicholson and T. P. Husband, "DIURNAL BEHAVIOR OF THE AGRIMI, CAPRA-AEGAGRUS," Journal of Mammalogy, vol. 73, pp. 135-142, Feb 1992.