Data Min Knowl Disc
DOI 10.1007/s10618-011-0227-9

# Mining periodic behaviors of object movements for animal and biological sustainability studies

**Zhenhui Li · Jiawei Han · Bolin Ding · Roland Kays**

Received: 25 May 2010 / Accepted: 28 May 2011


**Abstract** Periodicity is one of the most frequently occurring phenomena for moving objects. Animals usually have periodic movement behaviors, such as daily foraging behaviors or yearly migration behaviors. Such periodic behaviors are the keys to understand animal movement and they also reflect the seasonal, climate, or environmental changes of the ecosystem. However, periodic behaviors could be complicated, involving multiple interleaving periods, partial time span, and spatiotemporal noises and outliers. In this paper, we address the problem of mining periodic behaviors for moving objects. It involves two sub-problems: *how to detect the periods in complex movements*, and *how to mine periodic behaviors*. A period is usually a single value, such as 24 h. And a periodic behavior is a statistical description of the periodic movement for one specific period. For example, we could describe an animal's daily behavior in the way that "From 6 pm to 6 am, it has 90% probability staying at location $A$ and from 7 am to 5 pm, it has 70% probability staying at location $B$ and 30% probability staying at location $C$". So our tasks is to first detect the periods and

Responsible editor: Katharina Morik, Kanishka Bhaduri and Hillol Kargupta.

This is an extended version of our conference paper (Li et al. 2010b).

Z. Li (✉) · J. Han · B. Ding
University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: zli28@uiuc.edu

J. Han
e-mail: hanj@uiuc.edu

B. Ding
e-mail: bding3@uiuc.edu

R. Kays
New York State Museum, Albany, NY, USA
e-mail: rkays@mail.nysed.gov

Published online: 12 June 2011

then describe each periodic behavior according to different periods. Our main assumption is that the observed movement is generated from multiple interleaved *periodic behaviors* associated with certain *reference locations*. Based on this assumption, we propose a two-stage algorithm, Periodica, to solve the problem. At the first stage, the notion of *reference spot* is proposed to capture the reference locations. Through reference spots, multiple periods in the movement can be retrieved using a method that combines Fourier transform and autocorrelation. At the second stage, a *probabilistic model* is proposed to characterize the periodic behaviors. For a specific period, periodic behaviors are statistically generalized from partial movement sequences through hierarchical clustering. Finally, we show two extensions to the Periodica algorithm: (1) missing data interpolation, and (2) future movement prediction. Empirical studies on both synthetic and real data sets demonstrate the effectiveness of the proposed method.

**Keywords** Data mining · Object movements · Periodicity · Pattern analysis · Animal and environmental studies

## 1 Introduction

With the increasing interests in ecological and environmental studies based on animal migration and movements (Sugden et al. 2006; Getz and Saltz 2008; Nathan et al. 2008), technologies haven been developed to efficiently track animals (Cooke et al. 2004), and statistical methods have been developed for analyzing movement data (Dalziel et al. 2008; Patternson et al. 2008; Wittemyer et al. 2008). In recent years, the fast development of positioning technology (GPS) makes animal tracking easier, with higher resolution and longer duration. For example, MoveBank[1] is an organization for biologists to share their animal movement data. They now have more than a hundred of animal movement datasets, including birds, raptors, and herbivores. It is interesting yet challenging for animal scientists to mine spatiotemporal patterns from the movement data. The movement patterns are the natural reflections of ecosystems, such as the quantity and quality of forage (Polis et al. 1997), nutrient distribution and cycles (McNaughton et al. 1997), intra- and inter-specific disease transmission (Cross et al. 2005), and the distribution and population dynamics of animals (McNaughton et al. 1985).

In this work, we focus on one of the most basic movement patterns, *periodic behaviors*. For example, large herbivores may use spatial memory to locate preferred food patches and return to high quality foraging locations (Hewitson et al. 2005). Returns to previously grazed areas may be a useful foraging strategy for large herbivores to consume regrowing vegetation in its high primary productivity stage (McNaughton et al. 1985). Moreover, these returns may accelerate nutrient cycling in highly grazed sites (McNaughton et al. 1997). Another example is bird migration. Golden eagles start migrating to South America in late October and go back to Alaska around mid March (McIntyre and Adams 1999). The selection of migration time and locations

[1] www.MoveBank.org.

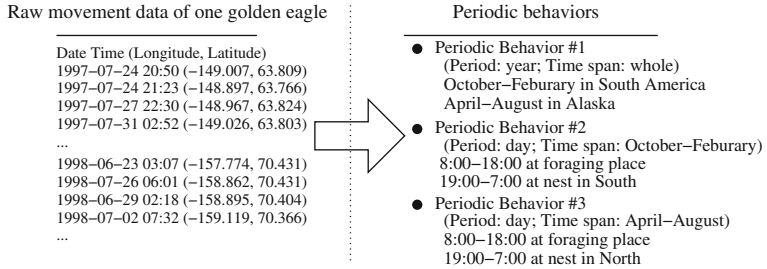


**Fig. 1** Periodic movement behaviors

is a reflection of temperature change and foraging quality. Such repeating behaviors in animal movement are referred to as "periodic behaviors". Discovery of periodic behaviors should contribute to our understanding of the habitat requirements of animals, the factors governing their space-use patterns, and interaction with the ecosystem (Bar-David et al. 2009). This understanding could lead to strategies for conservation and management of the animal populations and landscapes of interest. And thus, it should ultimately benefit the sustainability of an ecosystem.

However, mining periodic movement behaviors from long and noisy history data of a moving object is a challenging problem. For example, Fig. 1 shows the raw movement data of a golden eagle and the expected periodic behaviors. Based on manual examination of the raw data (on the left), it is almost impossible to extract the periodic behaviors (on the right). In fact, the periodic behaviors are quite complicated. There are multiple periods and periodic behaviors that may interleave with each other. Mining periodic behaviors can bridge the gap between raw data and semantic understanding of the data. The mining task includes two major issues.

First, the *periods* (i.e., the regular time intervals in a periodic behavior) are usually unknown. Even though there are many period detection techniques proposed in signal processing area, such as Fourier transform and autocorrelation, these methods cannot be *directly* applied to the spatiotemporal data. This is because the moving object will not repeat the movement by re-appearing at *exactly* the same point (in terms of $(x, y)$) on *exactly* the same time instance of a period. Besides, there could be *multiple* periods existing at the same time, such as the golden eagle may have one period as "day" and another as "year". If we consider the movement sequence as a whole, the longer period (i.e., year) will have fewer repeating times than the shorter period (i.e., day). So it is hard to select a threshold to find all the periods. Surprisingly, there is no previous work that can handle the issue about how to detect multiple periods from the noisy moving object data. To the best of our knowledge, there is only one work (Bar-David et al. 2009) that addresses the detection of periods for moving objects. It directly applies the Fourier transform on moving object data by transforming a location onto a complex plane. However, as shown in the toy example in Sect. 4, this method does not work in the presence of spatial noise.

Second, even if the periods are known, the *periodic behaviors* still need to be mined from the data because there could be *several* periodic behaviors with the same period. As one can see that, in golden eagle's movement, the same *period* (i.e., day)

is associated with two different *periodic behaviors*, one in the South and another in the North. In previous work, Mamoulis et al. (2004) studied the frequent periodic pattern mining problem for a moving object with a *given* period. However, the rigid definition of frequent periodic pattern does not encode the *statistical information*. It cannot describe the case such as "The eagle has 0.8 probability to be inside the nest at 6:00 everyday." One may argue that these frequent periodic patterns can be further summarized using probabilistic modeling approach (Yan et al. 2005; Wang and Parthasarathy 2006). But such models built on frequent periodic patterns do not truly reflect the real underlying periodic behaviors from the original movement, because frequent patterns are already a lossy summarization over the original data. Furthermore, if one can directly mine periodic behaviors on the original movement using polynomial time complexity, it is unnecessary to mine frequent periodic patterns and then summarize over these patterns.

In this paper, we formulate the periodic behavior mining problem and propose the assumption that the observed movement is generated from several *periodic behaviors* associated with some *reference locations*. We design a two-stage algorithm, Periodica, to detect the periods and further find the periodic behaviors.

At the first stage, we focus on detecting all the periods in the movement. Given the raw data as shown in Fig. 1, we use the kernel method to discover those reference locations, namely *reference spots*. The finding of reference spots is motivated by the idea of home range in biological study (Worton et al. 1989). That is, animals usually have their own home ranges when they move into a new region and repeat their activities at similar locations because of the seasonal foraging environment. Then, for each reference spot, the movement data is transformed from a spatial sequence to a binary sequence, which facilitates the detection of periods by filtering out the spatial noise. Besides, based on our assumption, every period will be associated with at least one reference spot. *All* periods in the movement can be detected if we try to detect the periods in every reference spot. At the second stage, we statistically model the periodic behavior using a *generative model*. Based on this model, underlying periodic behaviors are generalized from the movement using a hierarchical clustering method and the number of periodic behaviors is automatically detected by measuring the *representation error*.

Furthermore, we will examine two important extensions of Periodica in the study of periodic behaviors: (1) missing data interpolation, and (2) future data prediction. Because periodic behaviors provide us with the regularities in animal movements. Such regularity could be used to guess missing points in the data and also used to predict future movement. The raw data obtained from tracking facilities are usually transmitted with inconstant time gap. As we shown in Fig. 1, the golden eagle's movement could be recorded every several hours or every several days. But in most of real applications, people assume the data is sampled at constant rate by first linearly interpolating raw data. However, linear interpolation could introduce a lot of errors because the movement may not necessarily follow a linear model. For example, if the two consecutive recorded points for an eagle are (May 10th, 7:00 am, nest) and (May 15th, 8:00 am, nest). Linear interpolation will guess all the points from May 10th to May 15th are the nest, which may not be true. But if we already know the daily periodic behavior, we can better interpolate the missing data, such as guessing that the eagle

could be at the foraging place at 2:00 pm on May 11th. Similarly, periodic behaviors could also improve the performance of future movement prediction, especially for a distant querying time, such as one month after.

In summary, our major contributions are outlined as follows.

- We address an important problem in understanding movement data and formulate this problem as mining periodic behaviors.
- We propose algorithm Periodica to mine periodic behaviors, where Periodica is designed in the following two stages.
- We design a location-based method to effectively detect multiple periods in the movement using the concept of reference spots.
- We statistically model the periodic behavior, by proposing a clustering method, which determines the number of behaviors and discovers periodic behaviors.
- We examine two extensions of Periodica for further study of periodic behaviors: missing data interpolation and future data prediction.
- Comprehensive experiments are conducted on both real data and synthetic data, and the results demonstrate the effectiveness of our method.

The remaining of the paper is organized as follows. We discuss related work in Sect. 2. Section 3 formally states the problem and outlines the general framework. Section 4 introduces how to detect periods (stage 1). Section 5 describes the method to discover the periodic behaviors (stage 2). Two extensions of the method for the study of periodic behaviors are introduced in Sect. 6. We report our experimental results in Sect. 7 and conclude our study in Sect. 8.

## 2 Related work

### 2.1 Related work in computer science literature

A number of *periodic pattern mining* techniques have been proposed in data mining literature. But all the works are based on the assumption that *the periods are already given in advance*.

#### *2.1.1 Frequent periodic pattern mining*

Han et al. (1998, 1999) propose the algorithms for mining frequent partial periodic patterns. In their problem setting, each timestamp corresponds to a set of items. Different from previous works, Han et al's work (1998, 1999) considers partial periodicity, which is very common in practice since it is more likely that only some of the time episodes may exhibit periodic patterns. The goal is to find the partial frequent patterns that appear at least $min\_sup$ times. They present several algorithms for efficient mining of partial periodic patterns, by exploring some interesting properties related to partial periodicity, such as the Apriori property and the max-subpattern hit set property, and by shared mining of multiple periods.

Yang et al. (2000, 2004, 2002) and Wang et al. (2001) propose a series of works dealing with variations of periodic pattern mining, such as asynchronous patterns

(Yang et al. 2000), surprising periodic patterns (Yang et al. 2004), patterns with gap penalties (Yang et al. 2002), and higher level patterns (Wang et al. 2001). Asynchronous patterns (Yang et al. 2000) are the periodic patterns that may present only within a subsequence and whose occurrences may be shifted due to disturbance. Yang et al. (2004) introduces surprising periodic patterns, which is motivated by application in computational biology. That is, an infrequent pattern is considered very significant if its actual occurrence frequency exceeds the prior expectation by a large margin. They introduce a measurement, *information*, to value the degree of surprise of each occurrence of a pattern as a continuous and monotonically decreasing function of its probability of occurrence. Yang et al. (2002) is an extended work of Yang et al. (2004) which introduces the gap penalties. This work is also motivated from bio-informatics. They find that it is important to identify subseuqneces that a patterns repeats perfectly (or near perfectly). As a result, they extend the information gain measure in Yang et al. (2004) to include a penalty for gaps between pattern occurrences, named as generalized information gain. Similarly, the problem with gap constraint is studied in Zhang et al. (2005). To enforce gap constraint, Zhang et al. (2005) requires the characters in a pattern $P$ should match subsequences $S$ of original sequence in such a way that the matching characters in $S$ are separated by gaps of more or less the same size. Wang et al. (2001) studies the patterns that can be hierarchical in nature, where a higher level pattern may consist of repetitions of lower level patterns.

There are many works on mining spatio-temporal patterns (Wang et al. 2003; Mamoulis et al. 2004; Cao et al. 2005; Li et al. 2010a). Mamoulis et al. (2004) detects the periodic patterns for moving objects. However, the work takes period as an input without discussing how to detect period automatically. Besides, frequent periodic patterns cannot capture the statistical information as the periodic behaviors. Similar to our definition of periodic behavior, Indyk et al. (2000) studies the problem of discovering the most representative trend that repeats itself every $T$ timestamps. However, they can only discover one trend for a given period $T$ and such trend covers the whole time span. More recently, Lahiri and Berger-Wolf (2008) studies periodic behavior mining in dynamic social networks. Their problem focuses on the graphs that change dynamically over time. They try to detect the frequent periodic subgraph with a given period.

### *2.1.2 Automatic period detection in time series*

There are also works addressing the automatic period detection problem (Indyk et al. 2000; Yang et al. 2000; Ma and Hellerstein 2001; Berberidis et al. 2002; Cao et al. 2004; Elfeky et al. 2005a,b). Ma and Hellerstein (2001) and Yang et al. (2000) have developed a similar linear distance-based algorithm for discovering the potential periods regarding the symbols of the time series. But this method misses some valid periods since it only considers the adjacent intervals. In Cao et al. (2004), a data structure, the abbreviated list table (ALT) is proposed to compute the periods and the pattern. But such period is based on the threshold of $min\_sup$ which is not appropriate in our problem. Indyk et al. (2000) develops an $O(n \log^2 n)$ time complexity algorithm using sketch approaches to find representative trend where $n$ is the length of sequence. But only one period is detected in the whole sequence. Berberidis et al. (2002) detects the

period candidates for each symbol using autocorrelation. Improved from Berberidis et al. (2002) and Elfeky et al. (2005a) proposes a more efficient convolution method which considers multiple symbols together while detecting the period. However, as addressed in Sect. 4.2, both autocorrelation and convolution will detect a large set of period candidates, most of which are redundant. In Elfeky et al. (2005b), a method based on time warping is proposed, which is robust in the presence of shifting noise but is less efficient with time complexity $O(n^3)$.

### 2.2 Related work in biological literature

Surprisingly, we do not find any related work in computer science literature that directly addresses the period detection problem for moving objects. There is one work (Bar-David et al. 2009) in biological literature that studies the path recursion with application to African buffalo in South Africa. The path recursions defined in Bar-David et al. (2009) is similar to our periodic behavior definition. Path recursions are defined as repeated visits to a particular site or patch. They think such recursion analyses can provide biologists with a basis for inferring aspects of the process governing the production of buffalo recursion patterns, particularly the potential influence of resource recovery rate. They give a comprehensive discussion of how recursion analyses can be used when appropriate ecological data are available to elucidate various factors influencing movement. These factors include various limiting and preferred resources, parasites, and topographical and landscape factor.

The core technique in Bar-David et al. (2009) to detect periods in the movement include two parts: (1) recursion analysis, which identifies all closed paths, their length and locations, based on the observation that along a path that closes on itself, the sum of vector displacements is zero; and (2) circle analysis, which uses complex Fourier transform to display the periodogram of clockwise and counterclockwise looping in the movement patterns. But as we will show in Sect. 4, such method could be sensitive to noise and they can hardly detect multiple periods existing in only partial movements.

## 3 Framework overview

Let $D = \{(x_1, y_1, time_1), (x_2, y_2, time_2), \ldots\}$ be the original movement database for a moving object. The raw data is linearly interpolated with constant time gap, such as hour or day. The constant time gap depends on the sampling rate of the raw data. If the data is collected about every hour, we could use hour to make the trajectory sequence evenly gapped. The interpolated sequence is denoted as $LOC = loc_1 loc_2 \ldots loc_n$, where $loc_i$ is a spatial point represented as a pair $(loc_i.x, loc_i.y)$.

Given a location sequence $LOC$, our problem aims at mining all periodic behaviors. Before defining periodic behavior, we first define some concepts. A *reference spot* is a dense area that is frequently visited in the movement. The set of all reference spots is denoted as $O = \{o_1, o_2, \ldots, o_d\}$, where $d$ is the number of reference spots. A *period* $T$ is a regular time interval in the (partial) movement. Let $t_i$ $(1 \leq i \leq T)$ denote the $i$-th *relative timestamp* in $T$.

**Table 1** A daily periodic behavior of a golden eagle

| | 8:00 | 9:00 | 10:00 | ··· | 17:00 | 18:00 | 19:00 |
|---|---|---|---|---|---|---|---|
| Nest | 0.9 | 0.2 | 0.1 | ··· | 0.2 | 0.7 | 0.8 |
| Foraging place | 0.05 | 0.7 | 0.95 | ··· | 0.75 | 0.2 | 0.1 |
| Unknown | 0.05 | 0.1 | 0.05 | ··· | 0.05 | 0.1 | 0.1 |

A *periodic behavior* can be represented as a pair $\langle T, \mathbf{P}\rangle$, where $\mathbf{P}$ is a probability distribution matrix. Each entry $\mathbf{P}_{i,k}(1 \leq i \leq d, 1 \leq k \leq T)$ of $\mathbf{P}$ is the probability that the moving object is at the reference spot $o_i$ at relative timestamp $t_k$. The formal statistical modeling of periodic behavior will be given in Sect. 5.1.

*Example 1* Suppose $T = 24$ (h). The golden eagle's daily periodic behavior (Fig. 1 that involves with 2 reference spots (i.e., "nest" and "foraging place") could be represented as $(2 + 1) \times 24$ probability distribution matrix, as shown in Table 1. This table is an intuitive explanation of formal output of periodic behaviors, which is not calculated according to specific data in Fig. 1. The probability matrix encodes the noises and uncertainties in the movement. It statistically characterizes the periodic behavior, such as "The golden eagle starts going out for foraging *around* 8:00 in the morning."

**Definition 1 (Periodic Behavior Mining)** Given a length-$n$ movement sequence $LOC$, our goal is to mine all the periodic behaviors $\{\langle T, \mathbf{P}\rangle\}$.

There are two subtasks in the periodic behavior mining problem, detecting the periods and mining the periodic behaviors. We propose a two-stage algorithm, Periodica, where the overall procedure of the algorithm is developed in two stages and each stage targets one subtask.

Algorithm 1 shows the general framework of Periodica. At the first stage, we first find all the reference spots (Line 2) and for each reference spot, the periods are detected

**Algorithm 1** Periodica

INPUT: A movement sequence $LOC = loc_1 loc_2 \ldots loc_n$.
OUTPUT: A set of periodic behaviors.
ALGORITHM:
1: /* Stage 1: Detect periods (Sect. 4)*/
2: Find reference spots $O = \{o_1, o_2, \ldots, o_d\}$;
3: **for** each $o_i \in O$ **do**
4: Detect periods in $o_i$ and store the periods in $P_i$;
5: $P_{set} \leftarrow P_{set} \cup P_i$;
6: **end for**
7: /* Stage 2: Mine periodic behaviors (Sect. 5) */
8: **for** each $T \in P_{set}$ **do**
9: $O_T = \{o_i | T \in P_i\}$;
10: Construct the symbolized sequence $S$ using $O_T$;
11: Mine periodic behaviors in $S$.
12: **end for**

(Lines 3–5). Then for every period $T$, we consider the reference spots with period $T$ and further mine the corresponding periodic behaviors (Lines 7–10).

## 4 Detecting period

In this section, we discuss how to detect periods in the movement data. This includes two subproblems, namely, finding reference spots and detecting periods on binary sequence generated by these spots. First of all, we want to show why the idea of reference spots is essential for period detection. Consider the following example.

*Example 2* We generate a movement dataset simulating an animal's daily activities. Every day, this animal has 8 h staying at the den and the rest time going to some random places hunting for food. Figure 2a shows its trajectories. We first try the method introduced in Bar-David et al. (2009). The method transforms locations $(x, y)$ onto complex plane and use Fourier transform to detect the periods. However, as shown in Figs. 2b, c, there is no strong signal corresponding to the correct period because such a method is sensitive to the spatial noise. If the object does not follow more or less the same hunting *route* every day, the period can hardly be detected. However, in real cases, few objects repeat the exactly same route in the periodic movement.

Our key observation is that, if we view the data from the den, the period is easier to be detected. In Fig. 2d, we transform the movement into a binary sequence, where 1 represents the animal is at den and 0 when it goes out. It is easy to see the regularity in this binary sequence. Our idea is to find some important reference locations, namely *reference spots*, to view the movement. In this example, the den serves as our reference spot.

The notion of reference spots has several merits. First, it *filters out the spatial noise* and transforms the period detection problem from a 2-dimensional space (i.e., spatial) to a 1-dimensional space (i.e., binary). As shown in Fig. 2d, we do not care where the animal goes when it is out of the den. As long as it follows a regular pattern going out and coming back to the den, there is a period associated with the den. Second, we can detect *multiple* periods in the movement. Consider the scenario that there is a daily period with one reference spot and a weekly period with another reference spot, it is possible that only period "day" is discovered because the shorter period will repeat more times. But if we view the movement from two reference spots separately, both periods can be individually detected. Third, based on the assumption that each periodic behavior is associated with some reference locations, all the periods can be found through reference spots.

The rest of this section will discuss in details how to find reference spots and detect the periods on the binary sequence for each reference spot.

### 4.1 Finding reference spots

Since an object with periodic movement will repeatedly visit some specific places, if we only consider the spatial information of the movement, reference spots are those

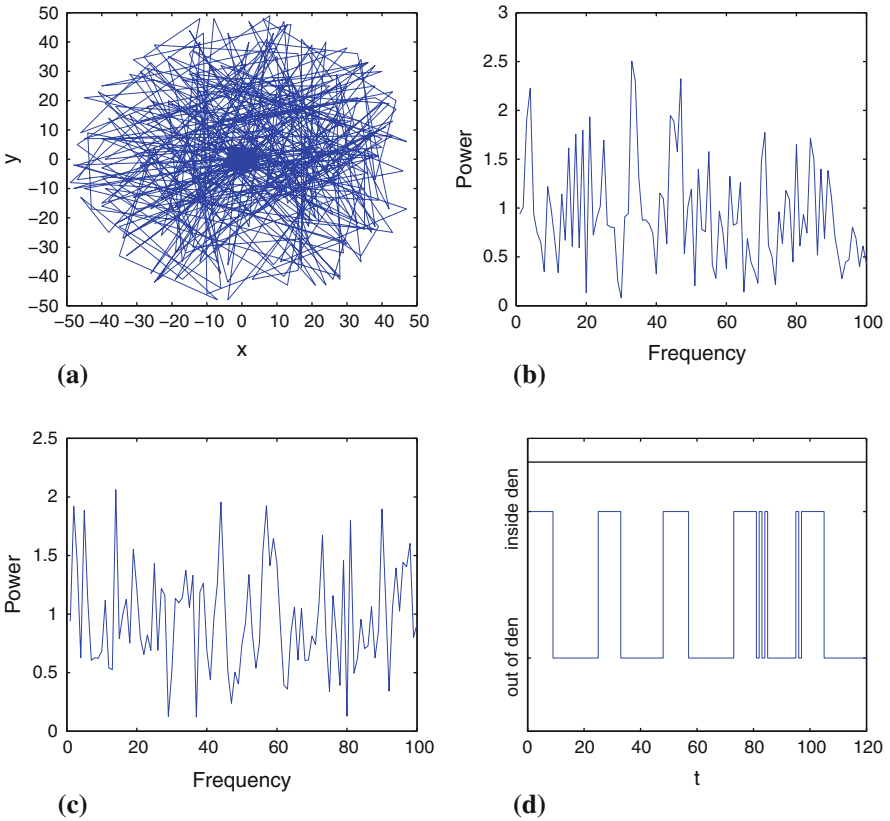


**Fig. 2** Illustration of the importance to view movement from reference spots. **a** Raw trajectories. **b** Fourier transform on $x + yi$. **c** Fourier transform on $y + xi$. **d** Binary sequence as viewed from the den

dense regions containing more points than the other regions. Note that the reference spots are obtained for individual object. While computing the density for each location in a continuous space is computationally expensive, we discretize the space into a regular $w \times h$ grid and compute the density for each cell. The grid size is determined by the desired resolution to view the spatial data.

To estimate the density of each cell, we adapt a popular kernel method (Worton et al. 1989), which is designed for the purpose of finding home ranges of animals. If an animal has frequent activities at one place, this place will have higher probability to be its home. This actually aligns very well with our definition of reference spots.

For each grid cell $c$, the density is estimated using the bivariate normal density kernel,

$$f(c) = \frac{1}{n\gamma^2} \sum_{i=1}^{n} \frac{1}{2\pi} \exp\left(-\frac{|c - loc_i|^2}{2\gamma^2}\right),$$

where $|c - loc_i|$ is the distance between cell $c$ and location $loc_i$. In addition, $\gamma$ is a smoothing parameter which is determined by the following heuristic method (Worton et al. 1989),

$$\gamma = \frac{1}{2}\left(\sigma_x^2 + \sigma_y^2\right)^{\frac{1}{2}} n^{-\frac{1}{6}},$$

where $\sigma_x$ and $\sigma_y$ are the standard deviations of the whole sequence $LOC$ in its $x$ and $y$-coordinates, respectively. The time complexity for this method is $O(whn)$.

After obtaining the density values, a reference spot can be defined by a contour line on the map. A contour line joins the cells of equal density. We use contour line to define the boundary of a reference spot. Any point within the reference spot has higher density value than that of the boundary. So the reference spot is essentially an area with high density.

The density value of a contour line can be determined as the top-$p\%$ density value among all the density values of all cells. The larger the value $p$ is, the bigger the size of reference spot is. In practice, $p$ can be chosen based on prior knowledge about the size of the reference spots. In many real applications, we can assume that the reference spots are usually very small on a large map (e.g. within 10% of whole area). So, by setting $p\% = 15\%$, most parts of reference spots should be detected with high probability. Even though it could introduce a small amount of additional noise at the same time, our period detection is robust in terms of noise as shown in experiment, specifically in Fig. 11.

*Example 3 (Running Example)* We will use a running example throughout the paper to illustrate our methods. Assume that a bird stays in a nest for half a year and moves to another nest staying for another half year. At each nest, it has a daily periodic behavior of going out for food during the daytime and coming back to the nest at night.

As shown in Fig. 3, the two small areas (spot #2 and spot #3) are the two nests and the bigger region is the food resource (spot #1). Figure 3a shows the density calculated using the kernel method. The grid size is $100 \times 100$. The darker the color is, the higher the density is. Figure 3b is the reference spots identified by contour using top-15% density value threshold.

### 4.2 Periods detection on binary sequence

Given a set of reference spots, we further propose a method to obtain the potential periods within *each* spot *separately*. Viewed from a single reference spot, the movement sequence now can be transformed into a binary sequence $B = b_1 b_2 \ldots b_n$, where $b_i = 1$ when this object is within the reference spot at timestamp $i$ and 0 otherwise. In discrete signal processing area, to detect periods in a sequence, the most popular methods are Fourier transform and autocorrelation, which essentially complement each other in the following sense, as discussed in Vlachos et al. (2005). On one hand, Fourier transform often suffers from the low resolution problem in the low frequency region, hence provides poor estimation of large periods. Also, the well-known spectral

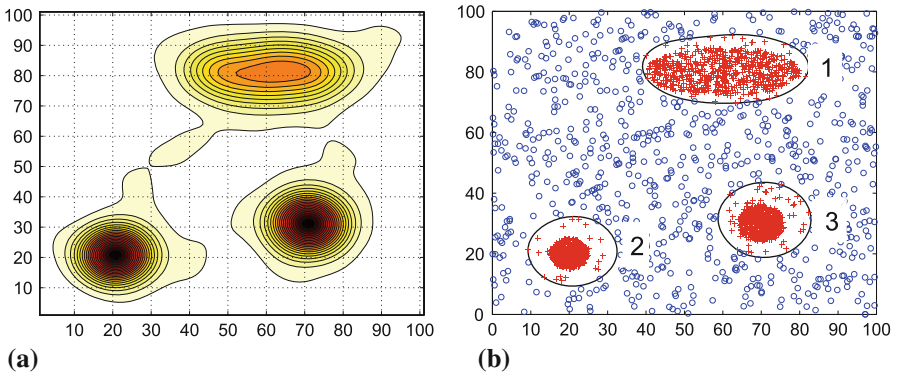


**Fig. 3** Finding reference spots. **a** Density map calculated by kernel method. **b** Reference spots defined by contours

leakage problem of Fourier transform tends to generate a lot of false positives in the periodogram. On the other hand, autocorrelation offers accurate estimation for both short and large periods, but is more difficult to set the significance threshold for important periods. Consequently, Vlachos et al. (2005) proposed to combine Fourier transform and autocorrelation to find periods. Here, we adapt this approach to find periods in the binary sequence $B$. Due to the space limit, we will briefly introduce the method. In order to get a more thorough understanding of the approach, we recommend readers to read (Vlachos et al. 2005).

In Discrete Fourier Transform (DFT), the sequence $B = b_1 b_2 \dots b_n$ is transformed into the sequence of $n$ complex numbers $X_1, X_2, \dots, X_n$. Given coefficients $X$, the periodogram is defined as the squared length of each Fourier coefficient: $F_k = \|X_k\|^2$. Here, $F_k$ is the power of frequency $k$. In order to specify which frequencies are important, we need to set a threshold and identify those higher frequencies than this threshold.

The threshold is determined using the following method. Let $B'$ be a randomly permutated sequence from $B$. Since $B'$ should not exhibit any periodicities, even the maximum power does not indicate the period in the sequence. Therefore, we record its maximum power as $p_{max}$, and only the frequencies in $B$ that have higher power than $p_{max}$ may correspond to real periods. To provide a 99% confidence level on what frequencies are important, we repeat the above random permutation experiment 100 times and record the maximum power of each permutated sequence. The 99-th largest value of these 100 experiments will serve as a good estimator of the power threshold.

Given that $F_k$ is larger than the power threshold, we still need to determine the exact period in the time domain, because a single value $k$ in *frequency domain* corresponds to a range of periods $[\frac{n}{k}, \frac{n}{k-1})$ in *time domain*. In order to do this, we use circular autocorrelation, which examines how similar a sequence is to its previous values for different $\tau$ lags: $R(\tau) = \sum_{i=1}^{n} b_\tau b_{i+\tau}$.

Thus, for each period range $[l, r)$ given by the periodogram, we test whether there is a peak in $\{R(l), R(l+1), \dots, R(r-1)\}$ by fitting the data with a quadratic function. If the resulting function is concave in the period range, which indicates the existence of

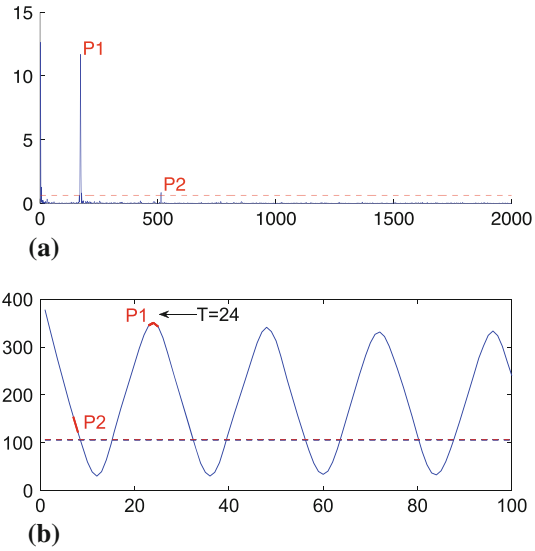


**Fig. 4** Finding periods. **a** Periodogram. **b** Circular autocorrelation

a peak, we return $t^* = \arg\max_{l \le t < r} R(t)$ as a detected period. Similarly, we employ a 99% confidence level to eliminate false positives caused by noise.

*Example 4 (Running Example (cont.))* The periodogram of reference spot #2 is shown in Fig. 4a. The red dashed line denotes the threshold of 99% confidence. There are two points $P_1$ and $P_2$ that are above the threshold. In Fig. 4b, $P_1$ and $P_2$ are mapped to a range of periods. We can see that there is only one peak, $P_1$, corresponding to $T = 24$ on the autocorrelation curve. This suggests the existence of a period of 1 day in the movement data.

Discrete Fourier Transform can be executed in $O(n \log n)$ time using Fast Fourier Transform algorithm (FFT). And since autocorrelation is a formal convolution which can also be solved by FFT, its complexity is also $O(n \log n)$. So, the overall time complexity of detecting periods in sequence $B$ is $O(n \log n)$.

## 5 Mining periodic behaviors

After obtaining the periods for each reference spot, now we study the task how to mine periodic behaviors. We will consider the reference spots with the same period together in order to obtain more concise and informative periodic behaviors. But, since a behavior may only exist in a *partial* movement, there could be several periodic behaviors with the same period. For example, there are two daily behaviors in a person's movement. One corresponds to the school days and the other one occurs during the summer. However, given a long history of movement and a period as a "day", we actually do not know how many periodic behaviors exist in this movement and which days belong to which periodic behavior. This motivates us to use a clustering method. Because the "days" that belong to the same periodic behavior should have the similar

temporal location pattern. We propose a generative model to measure the distance between two "days". Armed with such distance measure, we can further group the "days" into several clusters and each cluster represents one periodic behavior. As in the above example, "school days" should be grouped into one cluster and "summer days" should be grouped into another one.

In this section, we will formally present the technique to mine periodic behaviors. Since every period in the movement will be considered separately, *the rest of this section will focus on one specific period* $T$.

### 5.1 Modeling periodic behaviors

First, we retrieve all the reference spots with period $T$. By combining the reference spots with the same period together, we will get a more informative periodic behaviors associated with different reference spots. For example, we can summarize a student's daily behavior as "9:00–18:00 at office *and* 20:00–8:00 in the dorm". We do not consider combining two different periods in current work.

Let $O_T = \{o_1, o_2, \ldots, o_d\}$ denote reference spots with period $T$. For simplicity, we denote $o_0$ as any other locations outside the reference spots $o_1, o_2, \ldots, o_d$. Given $LOC = loc_1 loc_2 \ldots loc_n$, we generate the corresponding *symbolized movement sequence* $S = s_1 s_2 \ldots s_n$, where $s_i = j$ if $loc_i$ is within $o_j$. $S$ is further segmented into $m = \lfloor \frac{n}{T} \rfloor$ *segments*[2]. We use $I^j$ to denote the $j$-th segment and $t_k$ $(1 \leq k \leq T)$ to denote the $k$-th relative timestamp in a period. $I^j_k = i$ means that the object is within $o_i$ at $t_k$ in the $j$-th segment. For example, for $T = 24$ (h), a segment represents a "day", $t_9$ denotes 9:00 in a day, and $I^5_9 = 2$ means that the object is within $o_2$ at 9:00 in the 5-th day. Naturally, we may use the categorical distribution to model the probability of such events.

**Definition 2 (Categorical Distribution Matrix)** Let $\mathcal{T} = \{t_1, t_2, \ldots, t_T\}$ be a set of relative timestamps, $x_k$ be the categorical random variable indicating the selection of reference spot at timestamp $t_k$. $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_T]$ is a categorical distribution matrix with each column $\mathbf{p}_k = [p(x_k = 0), p(x_k = 1), \ldots, p(x_k = d)]^T$ being an independent categorical distribution vector satisfying $\sum_{i=0}^{d} p(x_k = i) = 1$.

Now, suppose $I^1$, $I^2$, ..., $I^l$ follow the same periodic behavior. The probability that the segment set $\mathcal{I} = \bigcup_{j=1}^{l} I^j$ is generated by some distribution matrix $\mathbf{P}$ is

$$P(\mathcal{I}|\mathbf{P}) = \prod_{I^j \in \mathcal{I}} \prod_{k=1}^{T} p(x_k = I^j_k).$$

According to maximum likelihood estimation (MLE), the best generative model can be defined as the optimal solution to the following log likelihood maximization problem:

[2] If $n$ is not a multiple of $T$, then the last $(n \mod T)$ positions are truncated.

$$\max_{\mathbf{P}} \left\{ L(\mathbf{P}|\mathcal{I}) = \log P(\mathcal{I}|\mathbf{P}) = \sum_{I^j \in \mathcal{I}} \sum_{k=1}^{T} p(x_k = I_k^j) \right\}. \tag{1}$$

The well-known solution to (1) is

$$p(x_k = i) = \frac{\sum_{I^j \in \mathcal{I}} \mathbf{1}_{I_k^j = i}}{|\mathcal{I}|}, \tag{2}$$

where $\mathbf{1}_A$ is the indicator function associated with the event $A$. That is, $p(x_k = i)$ is the relative frequency of reference spot $o_i$ at $t_k$ over all segments in $\mathcal{I}$.

Now, we formally define the concept of periodic behavior.

**Definition 3 (Periodic Behavior)** Let $\mathcal{I}$ be a set of segments. A periodic behavior over all the segments in $\mathcal{I}$, denoted as $\mathbf{H}(\mathcal{I})$, is a pair $\langle T, \mathbf{P} \rangle$. $T$ is the period and $\mathbf{P}$ is a probability distribution matrix learned through Eq. (2). We further let $|\mathcal{I}|$ denote the number of segments covered by this periodic behavior.

### 5.2 Discovery of periodic behaviors

With the definition of periodic behaviors, we are able to estimate periodic behaviors over a set of segments. Now given a set of segments $\{I^1, I^2, \ldots, I^m\}$, we need to discover which segments are generated by the same periodic behavior. Suppose there are $K$ underlying periodic behaviors, each of which exists in a partial movement, the segments should be partitioned into $K$ groups so that each group represents one periodic behavior.

A potential solution to this problem is to apply some clustering methods. In order to do this, a distance measure between two periodic behaviors needs to be defined. Since a behavior is represented as a pair $\langle T, \mathbf{P} \rangle$ and $T$ is fixed, the distance should be determined by their probability distribution matrices. Further, a small distance between two periodic behaviors should indicate that the segments contained in each behavior are likely to be generated from the same periodic behavior.

Several measures between the two probability distribution matrices $\mathbf{P}$ and $\mathbf{Q}$ can be used to fulfill these requirements. Here, since we assume the independence of variables across different timestamps, we propose to use the well-known Kullback-Leibler divergence as our distance measure:

$$KL(\mathbf{P}\|\mathbf{Q}) = \sum_{k=1}^{T} \sum_{i=0}^{d} p(x_k = i) \log \frac{p(x_k = i)}{q(x_k = i)}.$$

when $KL(\mathbf{P}\|\mathbf{Q})$ is small, it means that the two distribution matrices $\mathbf{P}$ and $\mathbf{Q}$ are similar, and vice versa.

Note that $KL(\mathbf{P}\|\mathbf{Q})$ becomes infinite when $p(x_k = i)$ or $q(x_k = i)$ has zero probability. To avoid this situation, we add to $p(x_k = i)$ (and $q(x_k = i)$) a background variable $u$ which is uniformly distributed among all reference spots,

$$p(x_k = i) = (1 - \lambda)p(x_k = i) + \lambda u, \tag{3}$$

where $\lambda$ is a small smoothing parameter $0 < \lambda < 1$. We usually set $\lambda = 0.01$.

To further understand from a statistical point of view why this is a good choice of distance measure for our problem, let us return to our generative model. Recall that $\mathcal{I}$ is the set of segments generated by $\mathbf{P}$, then $KL(\mathbf{P}\|\mathbf{Q})$ can be decomposed as

$$\begin{aligned}
KL(\mathbf{P}\|\mathbf{Q}) &= \sum_{k=1}^{T}\sum_{i=0}^{d} p(x_k = i)\log p(x_k = i) \\
&\quad - \sum_{k=1}^{T}\sum_{i=0}^{d} p(x_k = i)\log q(x_k = i) \\
&= -H(\mathbf{P}) - \sum_{k=1}^{T}\sum_{i=0}^{d} \frac{\sum_{I^j \in \mathcal{I}} \mathbf{1}_{I_k^j = i}}{|\mathcal{I}|}\log q(x_k = i) \\
&= -H(\mathbf{P}) - \frac{1}{|\mathcal{I}|}\sum_{I^j \in \mathcal{I}}\sum_{k=1}^{T}\log q(x_k = I_k^j) \\
&= -H(\mathbf{P}) - \frac{1}{|\mathcal{I}|}\log P(\mathcal{I}|\mathbf{Q}),
\end{aligned}$$

where $H(\mathbf{P})$ is the entropy of $\mathbf{P}$ and can be regarded as a constant in our problem. Thus, the KL-divergence measures how likely the segment set $\mathcal{I}$ can be generated by the distribution matrix $\mathbf{Q}$. In our clustering algorithm, among all possible choices of $\mathbf{Q}$, we simply select the one that maximizes the likelihood $P(\mathcal{I}|\mathbf{Q})$.

Now, suppose we have two periodic behaviors, $\mathbf{H}_1 = \langle T, \mathbf{P}\rangle$ and $\mathbf{H}_2 = \langle T, \mathbf{Q}\rangle$. We define the distance between these two behaviors as

$$dist(\mathbf{H}_1, \mathbf{H}_2) = KL(\mathbf{P}\|\mathbf{Q}).$$

Suppose there exist $K$ underlying periodic behaviors, there are many ways to group the segments into $K$ clusters with the distance measure defined. However, the number of underlying periodic behaviors (i.e., $K$) is usually unknown. So we propose a hierarchical agglomerative clustering method to group the segments while at the same time determine the optimal number of periodic behaviors. At each iteration of the hierarchical clustering, two clusters with the minimum distance are merged. We use a *representation error* to monitor the cluster quality. When the number of clusters turns from $k$ to $k - 1$, if the representation error increases dramatically, this indicates that $k$ could be the correct number of periodic behaviors. We will first describe the clustering method as Algorithm 2 assuming $K$ is given. The method to select optimal $K$ is introduced in Sect. 5.3.

Algorithm 2 illustrates the hierarchical clustering method. It starts with $m$ clusters (Line 1). A cluster $C$ is defined as a collection of segments. At each iteration, two clusters with the minimum distance are merged (Line 4–8). When two clusters are merged, the new cluster inherits the segments that owned by the original clusters $C_s$

**Algorithm 2** Mining periodic behaviors

INPUT: symbolized sequence $S$, period $T$, number of clusters $K$.
OUTPUT: $K$ periodic behaviors.
ALGORITHM:

1: segment $S$ into $m$ segments;
2: initialize $k = m$ clusters, each of which has one segment;
3: compute the pairwise distances among $C_1, \ldots, C_k$, $d_{ij} = dist(\mathbf{H}(C_i), \mathbf{H}(C_j))$;
4: **while** $(k > K)$ **do**
5: select $d_{st}$ such that $s, t = \arg\min_{i,j} d_{ij}$;
6: merge clusters $C_s$ and $C_t$ to a new cluster $C$;
7: calculate the distances between $C$ and the remaining clusters;
8: $k = k - 1$;
9: **end while**
10: return $\{\mathbf{H}(C_i), 1 \leq i \leq K\}$.

and $C_t$. It has a newly built behavior $\mathbf{H}(C) = \langle T, \mathbf{P} \rangle$ over the merged segments, where $\mathbf{P}$ is computed by the following updating rule:

$$\mathbf{P} = \frac{|C_s|}{|C_s| + |C_t|}\mathbf{P}_s + \frac{|C_t|}{|C_s| + |C_t|}\mathbf{P}_t. \tag{4}$$

Finally, $K$ periodic behaviors are returned (Line 9).

It takes $O(Td)$ to compute the distance between two behaviors, where $d$ is the number of reference spots. The number of iterations is $O(m)$. At each iteration, it takes $O(m \log m)$ to find the minimum pair and $O(mTd)$ to compute the distances between the newly merged cluster with other clusters. In summary, the complexity of the clustering algorithm is $O(m(mTd + m \log m)) = O(m^2 Td + m^2 \log m)$.

*Example 5 (Running Example (cont.))* There are two periodic behaviors with period $T = 24$ (h) in the bird's movement. Figure 5 shows the probability distribution matrix for each discovered periodic behavior. A close look at Fig. 5a shows that at time 0:00–8:00 and 22:00–24:00, the bird has a high probability being at reference spot #2, which is a nest shown in Fig. 3b. At time 12:00–18:00, it is very likely to be at reference spot #1, which is the food resources shown in Fig. 3b. And at the time 9:00–11:00, there are also some probability that the bird is at reference spot #1 or reference spot #2. This indicates the bird goes out of the nest around 8:00 and arrives at the food resources place around 12:00. Such periodic behaviors well represent the bird's movement and truly reveal the mechanism we employed to generate this synthetic data.

### 5.3 Number of periodic behaviors

In the clustering algorithm, $K$ represents the number of periodic behaviors in the movement sequence. Since it is unknown how many periodic behaviors are in the movement, it is important to find the right way to pick the appropriate parameter $K$.

Ideally, during the hierarchical agglomerative clustering, the segments generated from the same behavior should be merged first because they have smaller KL-divergence distance. Thus, we judge a cluster is good if all the segments in the cluster are

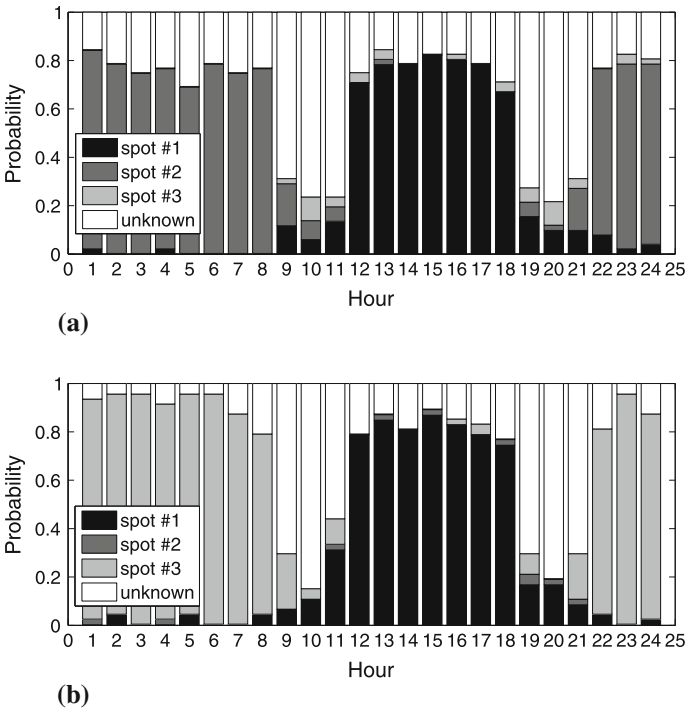


**Fig. 5** Periodic behaviors. **a P** of periodic behavior #1. **b P** of periodic behavior #2

concentrated in one single reference spot at a particular timestamp. Hence, a natural representation error measure to evaluate the representation quality of a cluster is as follows. Note that here we exclude the reference spot $o_0$ which essentially means the location is unknown.

**Definition 4 (Representation Error)** Given a set of segments $C = \{I^1, I^2, \ldots, I^l\}$ and its periodic behavior $\mathbf{H}(C) = \langle T, \mathbf{P} \rangle$, the representation error is,

$$E(C) = \frac{\sum_{I^j \in C} \sum_{i=1}^{T} \mathbf{1}_{I_i^j \neq 0} (1 - p(x_i = I_i^j))}{\sum_{I^j \in C} \sum_{i=1}^{T} \mathbf{1}_{I_i^j \neq 0}}.$$

At each iteration, all the segments are partitioned into $k$ clusters $\{C_1, C_2, \ldots, C_k\}$. The overall representation error at current iteration is calculated as the mean over all clusters,

$$\mathcal{E}_k = \frac{1}{k} \sum_{i=1}^{k} E(C_i).$$

During the clustering process, we monitor the change of $\mathcal{E}_k$. If $\mathcal{E}_k$ exhibits a dramatical increases comparing with $\mathcal{E}_{k-1}$, it is a sign the newly merged cluster may contain two

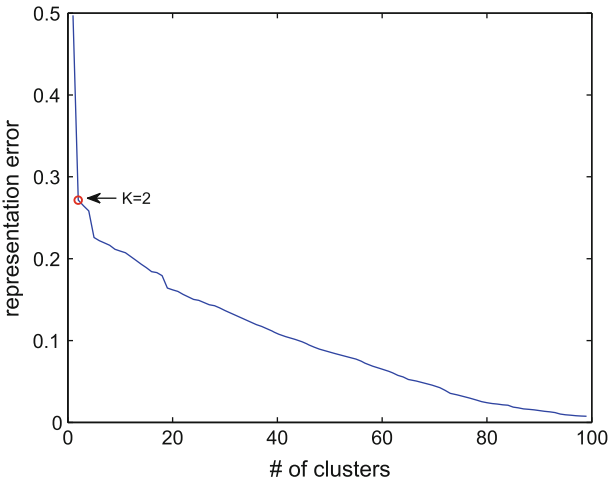


**Fig. 6** Representation error

different behaviors and $k-1$ is likely to be a good choice of $K$. The degree of such change can be observed from the derivative of $\mathcal{E}$ over $k$, $\frac{\partial \mathcal{E}}{\partial k}$. Since a sudden increase of $\mathcal{E}$ will result in a peak in its derivative, we can find the optimal $K$ as $K = \arg\max_k \frac{\partial \mathcal{E}}{\partial k}$.

*Example 6 (Running Example (cont.))* As we can see Fig. 6, the representation error suddenly increases at $k = 2$. This indicates that there are actually two periodic behaviors in the movement. This is true because the bird has one daily periodic behavior at the first nest and later has another one at the second nest.

## 6 Extensions

In this section, we extend our work in two directions by using periodic behaviors mined from the movement. One is missing data interpolation, that is to use periodic behaviors to estimate previous missing points in the movement. Another is prediction for future movement. If we assume future movement also complies with some periodic behavior, we could get a better prediction of its future movement. We will discuss in more details of the data interpolation and we will also conduct experiment on this in Sect. 7.4. Prediction is a more complicated topic with stronger assumption on future movement. So we will only briefly introduce the idea how to make of use of periodic behaviors for better prediction and leave the details of this method as one future work.

### 6.1 Missing data interpolation

The movement data obtained from most tracking devices are not recorded at the constant rate. For example, due to battery limit, the time gaps between two consecutive locations in golden eagles movement could possibly be several days or several min-

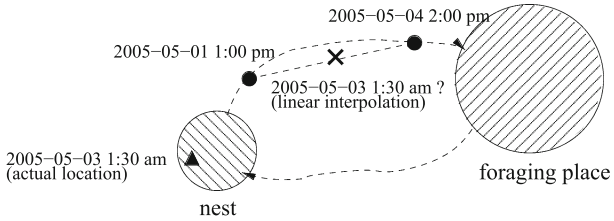


**Fig. 7** Incorrect estimation of missing data by linear interpolation

utes. And from our examination of the taxi data, the movement could have really high-resolution (e.g., a recorded location point in every few seconds). But when people turn off the tracking device in the car, there could be hours or even days of missing data for this vehicle.

However, in most problems, people usually assume the given data are recorded at constant rate (e.g., every minute or every hour). So it is essential to pre-process the data to make it constantly sampled. A simple and straightforward method is to use linear interpolation on the missing points. Given two consecutive recorded locations, $(loc_1, t_1)$ and $(loc_2, t_2)$, and an expected recording time $t$ $(t_1 < t < t_2)$, the missing point at time $t$ is linearly interpolated as:

$$loc_1 + (loc_2 - loc_1) \times \frac{t - t_1}{t_2 - t_1}.$$

Linear interpolation is easy to implement and it is suitable for the case when the missing time period is considerably short. But when the data is sparse and there are long time periods that the data is missing, such linear interpolation could introduce a lot of errors on the estimation of real movement. For example, in Fig. 7, the bird could have a daily periodic pattern from its nest to the foraging place. Assume the two recorded timestamps are 2005-05-01 1:00 pm and 2005-05-04 2:00 pm. By linear interpolation, the estimated location at 2005-05-03 1:30 am should be right in the middle of two recorded locations. However, if we already know the periodic behavior of this bird, we could infer that this bird should stay at nest at night. Taking Fig. 7 as an example, if we already know the bird has a daily periodic behavior and every day in May, 2005 belong to the same periodic behavior. Now, if we want to estimate the location at 2005-05-03 1:30 am, we can use the locations at 1:30 am in other days to guess its location. If we find out that most of the locations at 1:30 am for other days are inside the nest, we can guess that it was also at nest on 2005-05-03 1:30 am. Such interpolation using periodic behavior could be more accurate than linear interpolation.

We now introduce the method of interpolation missing data using periodic behaviors. As mentioned in Sect. 3, the raw data is first linearly interpolated with constant time gap and the interpolated sequence is $LOC = loc_1 loc_2 \ldots loc_n$. Let $loc_x (1 \leq x \leq n)$ be an estimated location and now we want to estimate its actual location using periodic behaviors. Suppose $loc_x$ belong to periodic behavior $\mathbf{H} =< T, \mathbf{P} >$. Note that here we only consider that case that $loc_x$ belongs to only one single periodic behavior. In the process of summarizing periodic behaviors, we know that there are a set of segments belong to this periodic behavior $\mathbf{H}$. These segments should exhibit

similar periodic behavior. We use $C = \{I^1, I^2, \ldots, I^l\}$ to denote these set of segments. Assume that $loc_x$ is a missing point in segment $I^y$. The estimated $loc_x$ can be computed as:

$$loc_x = \frac{\sum_{i \in [1,l], i \neq y} I^i_{x \bmod T}}{l-1}.$$

We want to give a short discussion on the pros and cons of two interpolation methods. Each interpolation method has its own merits when facing difference cases. Linear interpolation is simple and more general, especially for movements that do not have periodic behaviors. Even in our pre-processing step, we use linear interpolation to get a rough estimation of the movement. Linear interpolation is more suitable when the data is not sparse (e.g., the gap between two consecutive timestamps is in a hour) or the moving object is in the moving mode (because most of the moving objects would choose the shortest path to get to the destination). And interpolation using periodic behavior is more suitable when there is a long time duration of missing data (e.g., a few days). It is especially useful when the moving object has a major change in its locations in this long missing time period. For example, if two month data are missing in the movement and the bird moved from South to North at that time, the linear interpolation will assume the bird was always on the way from South to North in these two months. However, the migration may only take several days.

### 6.2 Prediction for future movement

Given historical movement data, it is useful to forecast future movement. In biological study, such movement prediction could help us protect endangered species. More interestingly, if the animals deviate from the expected route a lot, it could be a sign of ecological change. There have been many related works proposed to solve the prediction problem. Most existing techniques target at *near* future movement prediction, such as next minute or next hour. Linear motion functions (Saltenis et al. 2000; Tao and Papadias 2003; Tao et al. 2003; Jensen et al. 2004; Patel et al. 2004) have been extensively studied for movement prediction. More complicated models are studied in Tao et al. (2004). As pointed out by Jeung et al. (2008), the actual movement of a moving object may not necessarily comply with some mathematical models. It could be more complicated than what the mathematical formulas can represent. Moreover, such models built based on recent movement are not useful for predicting distant future movement, such as next day or one month after.

Periodic behaviors can help better predict future movement, especially for a distant query time. In Jeung et al. (2008), it proposes prediction method using periodic patterns. It assumes that the period $T$ and periodic patterns are already given. It builds an indexing structure, Trajectory Pattern Tree, which indexes the periodic patterns to answer predictive queries efficiently. Then, it proposes a Hybrid Prediction Algorithm that provides predictions for both near and distant time queries. For non-distant time queries, they use the Forward Query Processing which treats recent movements of an object as an important parameter to predict near future locations. A set of qualified

candidates will be retrieved and ranked by their premise similarities to the given query. Then they select top-$k$ patterns and return the centers of their consequences as answers. For a distant time queries, since recent movements become less important for prediction, the Backward Query Processing is used. Its main idea is to assign lower weights to premise similarity measure and higher weights to consequences which are closer to the query time in the ranking process of the pattern selection.

Work (Jeung et al. 2008) can be considered as an extension of using periodic behaviors. However, in Jeung et al. (2008), it assumes that there is only one period $T$. But in reality, there could be multiple periods interleaving with each other. For example, the birds could have yearly migration behavior and also daily foraging behavior. If we only use daily periodic behaviors for prediction, there could be many daily behaviors in many different places since birds might migrate to places very far away from each other. In such complicated cases, (Jeung et al. 2008) is likely to fail to predict the actual locations.

Therefore, it is important to identify which periodic behaviors that the object belongs to, which is also a challenging task. One possible approach is to iteratively refine the prediction. For example, when we want to predict a bird's future location at 10:00 am March 1st next year, we will first identify the periodic with lower time resolution, such as yearly migration behavior. By doing this, we may estimate the region that this bird could be in March. Then, we look into higher time resolution behaviors, such as daily behaviors that happened in this region. At 10:00 am, the bird may fly to foraging places to get the food. Then we can use the locations at 10:00 am in those daily behaviors to get a better estimation of the actual location. Such approach is only a tentative idea to solve the prediction in complicated real cases. We consider it as an interesting future work.

## 7 Experiment

In this section, we systematically evaluate the techniques presented in the paper. The language used is C++ and the experiments are performed on a 2.8 GHz Intel Core 2 Duo system with 4 GB memory. The system ran MAC OS X with version 10.5.5 and gcc 4.0.1.

### 7.1 Mining periodic behaviors

In this section, we test our periodic behavior mining algorithm, Periodica, on both real and synthetic data sets.

#### *7.1.1 Mining periodic behaviors on a real bald eagle data*

We test our method on a real dataset[3]. The data contains a 3-year tracking (2006.1–2008.12) of a bald eagle in the North America. The data is first linearly interpolated using the sampling rate as a day.

[3] The data set is obtained from www.movebank.org.

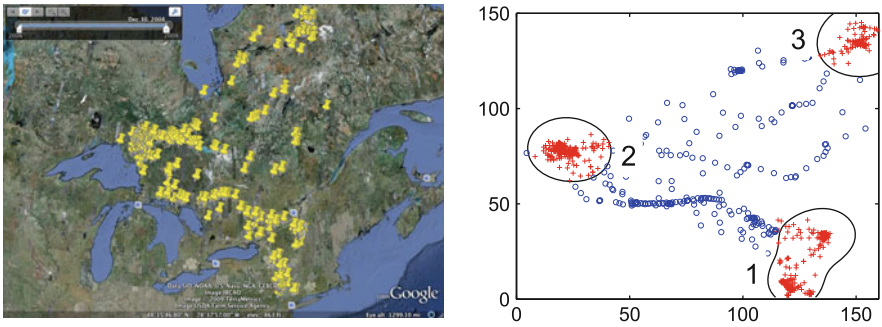


**Fig. 8** Real bald eagle data. **a** Raw data of bald eagle plotted on Google Earth. **b** Reference spots

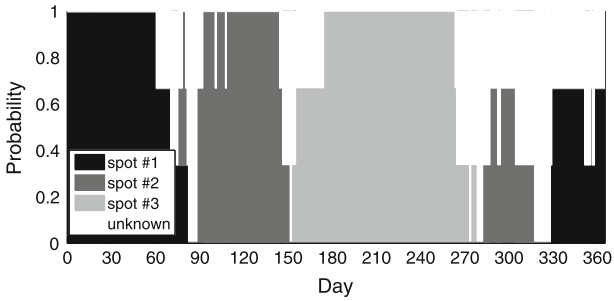


**Fig. 9** Periodic behaviors of bald eagle

Figure 8a shows the original data of bald eagle using Google Earth. It is an enlarged area of Northeast in America and Quebec area in Canada. As shown in Fig. 8b, three reference spots are detected in areas of New York, Great Lakes and Quebec. By applying period detection to each reference spot, we obtain the periods for each reference spots, which are 363, 363 and 364 days, respectively. The periods can be roughly explained as a year. It is a sign of yearly migration in the movement.

Now we check the periodic behaviors mined from the movement. Ideally, we want to consider three reference spots together because they all show yearly period. However, we may discover that the periods are not exactly the same for all the reference spots. This is a very practical issue. In real cases, we can hardly get perfectly the same period for some reference spots. So, we should relax our constraint and consider the reference spots with *similar* periods together. If the difference of periods is within some tolerance threshold, we take the average of these periods and set it as the common period. Here, we take period $T$ as 363 days, and the probability matrix is summarized in Fig. 9. Using such probability matrix, we can well explain the yearly migration behavior as follows.

"This bald eagle stays in New York area (i.e., reference spot # 1) from December to March. In March, it flies to Great Lakes area (i.e., reference spot #2) and stays there until the end of May. It flies to Quebec area (i.e., reference spot #3) in the summer and

stays there until late September. Then it flies back to Great Lake again staying there from mid October to mid November and goes back to New York in December."

This real example shows the periodic behaviors mined from the movement provides an insightful explanation for the movement data.

### *7.1.2 Mining periodic behaviors on synthetic data*

*Synthetic data generation.* In order to test the effectiveness under various scenarios, we design a generator for moving objects with periodicity according to a set of parameter values. These parameters are the length $n$ of the time history (in timestamps), period $T$, the probability $\alpha$ for a periodic segment in the object's movement to comply with regular movement, the probability $\beta$ for the noise for each timestamp in a regular periodic segment, and the variance $\sigma$ of normal distribution to add temporal perturbations to the periodic segment.

Before generating the movement, we first create several reference spots. Each reference spot is a small circle with radius ranges from 1% to 5% of the map size. A standard segment $seg_{std}$ with length $T$ is the movement following the regular periodic pattern. For example, for $T = 24$ (h), $seg_{std}$ could be designed as 6:00 pm–8:00 am at reference spot $A$ (such as home) and 8:30 am–5:30 pm h at reference spot $B$ (such as office). Then, the movement of the object is generated. For every segment $seg$, we first determine whether $s$ should be a regular segment or not, given the probability $\alpha$.

If $seg$ is a regular segment, the object's movement is generated as follows. According to standard segment, suppose that from timestamp $t_0$ to $t_1$ the object is at reference spot $A$, we further perturb $t_0$ and $t_1$ with some normal distribution (i.e., $t'_0 = N(t_0, \sigma^2)$, $t'_1 = N(t_1, \sigma^2)$). For all the experiments, we fix $\sigma = 0.5$. Finally, with probability $1 - \beta$, the object is at a random location within the circle of reference spot $A$ from $t'_0$ to $t'_1$. For other timestamps that are not confined to any reference spot, a random location is generated. If $seg$ is an irregular segment, for each timestamp, a random location is assigned.

*The case with multiple periods.* Since the running example has already illustrated periodic behaviors in partial movement, here we test our algorithm on a case with multiple periods. Suppose that there are 4 reference spots. Imagine them as "home", "office", "gym", and "class". A standard movement segment is generated as 20:00–8:00 at home every day; 9:00–14:00 at office on weekdays; 15:00–17:00 at gym on Tuesdays and Thursdays; 15:00–17:00 at class on Mondays, Wednesdays and Fridays. Furthermore, we choose $n = 8400$, $\alpha = 0.9$ and $\beta = 0.1$.

The periods detected for each reference spot are shown in Table 2. There are two periods detected: 24 (i.e., day) and 168 (i.e., week). It is interesting to see that office has both 24 and 168 as the periods. This is because office is visited "almost" every day except weekends. So both day and week are reasonable periods.

There is one daily behavior and one weekly behavior. Their probability matrices are illustrated in Fig. 10. In Fig. 10a, we can infer that this person leaves home *around* 8:00 am because the probability starts to drop at 8:00 am. In the weekly movement shown in Fig. 10b, 9:00–14:00 weekdays, the person stays in the office with high probability. Gym is involved with Tuesday and Thursday afternoons and class is

**Table 2** Periods detected

| Obs. spot | Home | Office | Gym | Class |
|---|---|---|---|---|
| Periods (h) | 24 | 24, 168 | 168 | 168 |

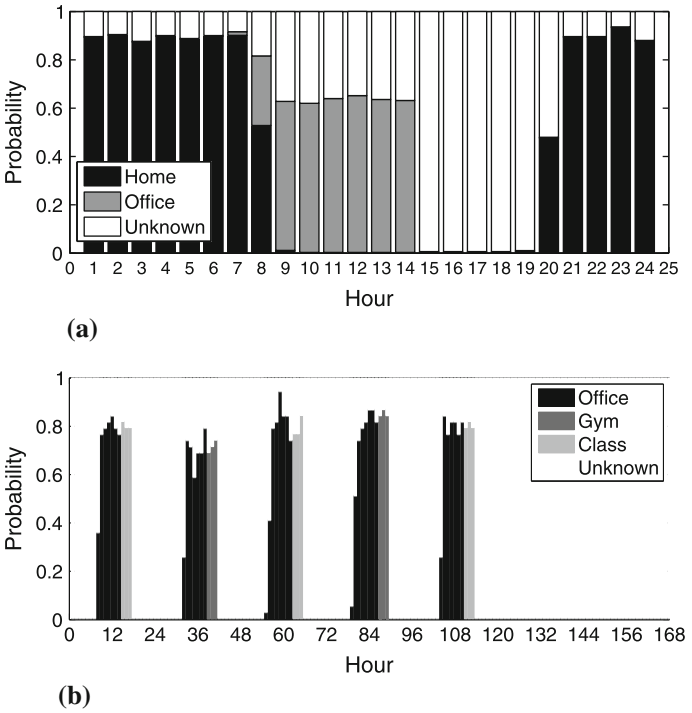


**Fig. 10** Periodic behaviors. **a** Periodic behavior for $T = 24$. **b** Periodic behavior for $T = 168$

involved with Monday, Wednesday and Friday afternoons. The behaviors on weekends are unknown.

*Performance evaluation w.r.t. different parameters in synthetic data generation.* We further verify the effectiveness of our algorithms with respect to the two parameters we introduced at the beginning of this section, $\alpha$ and $\beta$, on synthetic datasets. Recall that $\alpha$ represents the proportion of regular segments in the whole sequence and $\beta$ indicates the level of random noise. Again we use our *Running Example* to generate the synthetic data. This time, we vary $\alpha$ from 1 to 0.6, and simultaneously, we choose $\beta$ from 0 to 0.5. We test the effectiveness of the period detection algorithm and the summarization algorithm separately. All experiments are repeated 100 times and the results are averaged.

For the period detection algorithm, we report the success rates in Figure 11a. Since we know the ground truth ($T = 24$), we judge a trial is successful if among all detected periods, the one with the large correlation value is within the range [23, 25]. The result suggests that our period detection algorithm is nearly perfect in all cases with $\alpha \leq 0.8$.

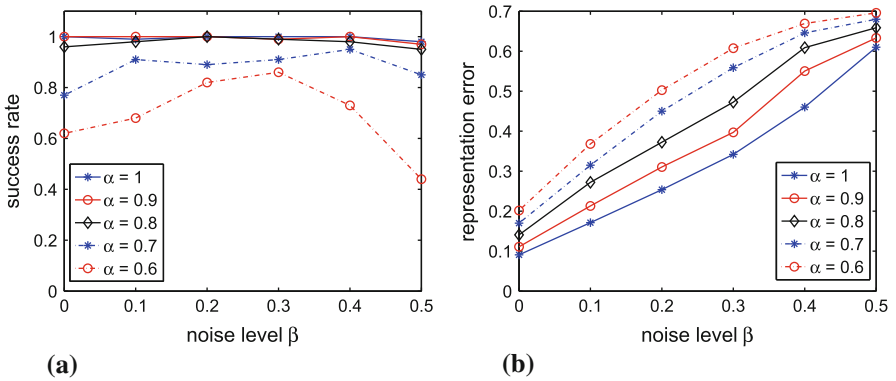


**Fig. 11** Performance evaluation. **a** Success rate of the period detection algorithm. **b** Representation error of the summarization algorithm

It is also noticeable that, compared to irregular segments, our algorithm is more robust to random noise, which may be caused by the failure of tracking devices or transmission networks during the data acquisition process. Furthermore, since irregular segments often reflects the changes of behaviors in the movement, the sensitivity to the irregular segments is also desirable for our algorithm which is designed for mining periodic behaviors.

For the summarization algorithm, we show in Figure 11b the representation error for $K = 10$ as defined in Sect. 5.3. To see the significance of the result, observe that, for example, with $\alpha = 0.9$ and $\beta = 0.1$, if we use 10 clusters to summarize all the daily segments of one year, the representation error is about 0.2. This means that we can obtain compact high-quality summarization even with moderate amount of irregularity and noise. This further shows that our algorithm is indeed able to filter out redundancy between the segments which are generated by periodic behaviors and therefore reveals the true behaviors.

### 7.2 Period detection comparison

Bar-David et al. (2009) applies Fourier Transform on spatio-temporal data to detect period in the movement. As a simple example shown in Sect. 4, such method is less resistant to noise in detecting periods. In this section, we will compare the method in Bar-David et al. (2009) with our Periodica to examine their robustness in terms of noise.

The synthetic data is generated as follows. Similar to Example 2 in Sect. 4, assume an animal has daily periodic movement. It has 8 h staying at the den and the rest time going to some random places hunting. Now we fix its den as the point (0,0) and foraging area as a $r \times r$ circle with center at (10,10). When the animal is out of den looking for food, it could appear at any random location within the circle. For example, Fig. 12 shows the synthetic movement with $r = 5$ and $r = 10$ individually. When $r$ gets larger, the noise in the movement increases.

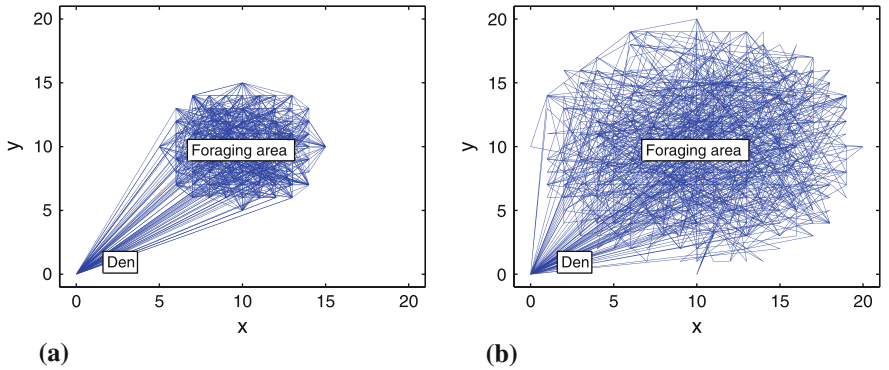


**Fig. 12** Synthetic data. **a** Synthetic data with small noise ($r = 5$). **b** Synthetic data with large noise ($r = 10$)

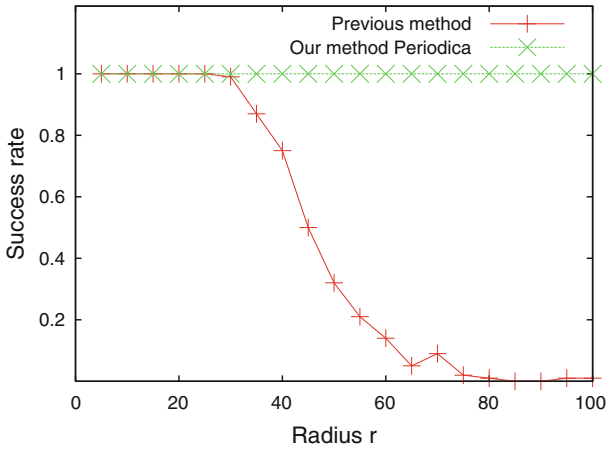


**Fig. 13** Success rate of Periodica compared with previous method (Bar-David et al. 2009)

Now we test previous method (Bar-David et al. 2009) and our method in terms of different radius sizes. We vary radius $r$ from 5 to 100. The synthetic data is generated with a given $r$. And the method is successful if it can detect the period as 24. For a fixed radius $r$, the test is repeated for 100 times and the success rate is the number of correct period detection among the 100 trials. Figure 13 shows the success rate of two methods. As we can see that Periodica is very robust in terms of noise. It can always detect the period correctly. However, previous method (Bar-David et al. 2009) fails quickly when the radius gets larger. This is because the method simply treats each location as a complex number. When the foraging area is getting big, the complex numbers are becoming more random in bigger amplitude and Fourier transform will be affected by that. But Periodica is not sensitive to that. Once the den is selected as the reference spot, the locations outside of den will be treated as the same.

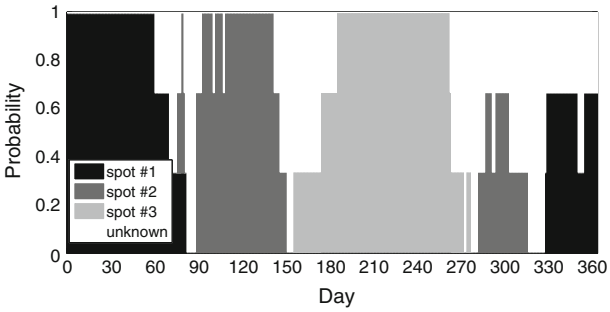


**Fig. 14** Periodic behavior of bald eagle with missing data

### 7.3 Missing data interpolation

In this section, we again use the bald eagle real data to demonstrate the effectiveness of interpolating missing points using periodic behaviors. The experiment setting is as follows. From the 3-year bald eagle tracking data, we manually remove part of them over the time span from May 15th to July 15th in the second year. This period corresponds to the time when the bald eagle migrate from reference spot 2 to reference spot 3 in Fig. 8b. This part of the data is then considered as missing and a linear interpolating is carried out before our period detection method is applied.

Our periodic behavior mining algorithm Periodica again detected the yearly migration behavior of the bald eagle and the mined periodic behavior is shown in Fig. 14.

Given the mined periodic behavior of the bald eagle, we can use the interpolation method proposed in Sect. 6 to estimate the missing data. For this particular case, it simply reduces to assign the average location of the corresponding timestamp in the first year and the third year to each missing entry in the second year.

The periodic interpolation result is compared with linear interpolation and the ground truth in Fig. 15. It clearly shows that periodic interpolation result agrees with the ground truth much better than linear interpolation. Moreover, the mean distance errors of periodic interpolation and linear interpolation are 22.6 and 43.4 on the map, respectively. That is, by exploring the periodic behavior of the subject, we reduces the interpolation error by almost a half.

In conclusion, we have shown that periodic interpolation is more accurate than linear interpolation for estimating missing data for real world moving objects, particularly over a long time span.

### 7.4 Prediction for future movement

In this section, we will examine the future movement prediction using periodic behavior. If a moving object has strong periodicity in its movement, we could use its historical movement to predict future locations. In this experiment setting, we will also use the bald eagle data. The data contains 3-year movement from 2006.1 to 2008.12. Assume

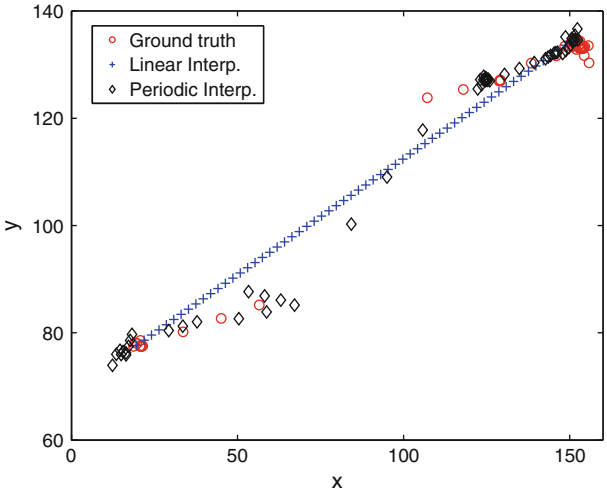


**Fig. 15** Comparison of missing data interpolation methods

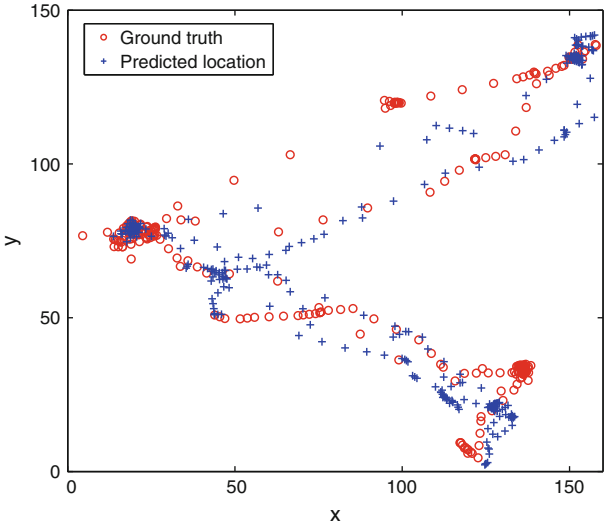


**Fig. 16** Predicted locations versus true locations in the third year of bald eagle movement

we already know that the period of this eagle is 363 days. Now taking the first 2-year data as the known movement, we want to predict the movement in the third year.

In Sect. 6.2, we discuss how to predict future movement using periodic behavior. In this case, we simply use the average location of the corresponding timestamps in the first year and the second year to predict the location of the third year. Fig. 16 shows the predicted locations and true locations in the third year. We can see that the overall trajectory of the third year can be predicted quite well. This is because the bald eagle has high periodicity in its movements.

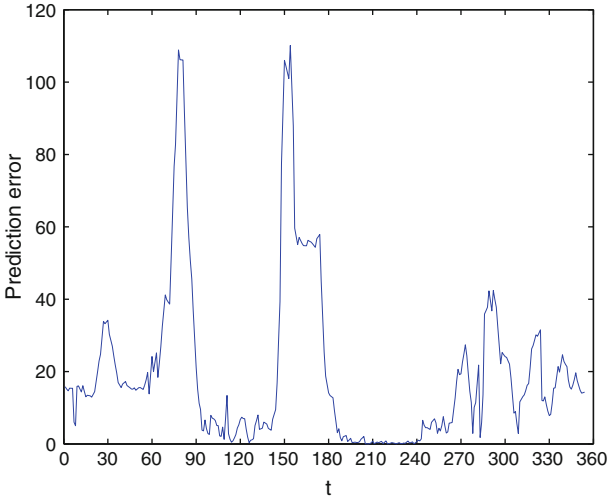


**Fig. 17** Prediction error for each day in the third year of bald eagle movement

Figure 17 depicts the distance between predicted location and true location for each day in the third year. As we discuss in Sect. 7.1.1, the eagle migrates in March, May, October and December. From Fig. 17, we can see that the prediction error during the migration time is relatively high, whereas the prediction error is really low when the eagle was staying at some location. The biggest advantage of using periodic behavior for prediction is that the prediction is not limited to near future. Previous methods (Saltenis et al. 2000; Tao and Papadias 2003; Tao et al. 2003; Jensen et al. 2004; Patel et al. 2004; Tao et al. 2004) using motion models usually predict locations in next few timestamps. But here we could even predict the movement in the next year as long as the eagle still follows such periodic behavior.

In conclusion, it is easy to use periodic behavior to predict future movement and the overall prediction is quite accurate. But we have only studied a simple case with one single period. When there are multiple interleaved periodic behaviors, it is more challenging to make accurate predictions.

## 8 Conclusion and future work

In this paper, we address an important and difficult problem: periodic behavior mining for moving objects. We propose a two-stage algorithm, Periodica. In the first stage, periods are detected through reference spots using Fourier transform and autocorrelation. In the second stage, periodic behaviors are statistically summarized using hierarchical clustering method. Empirically studies show that our method can deal with both noisy and complicated cases. A case study on a real data demonstrates the effectiveness of our method in practice. We further extend our work by discussing missing data interpolation and future movement prediction using periodic behaviors.

And the experiment on missing data interpolation shows that using periodic behaviors could better interpolate missing points.

While our approach fixes some reference spots using spatial information only, it is interesting to dynamically detect reference spots integrating with temporal information. This could give a more precise estimation on the reference locations. Another important issue is to find periodic behaviors in the data with the very sparse and inconstant sampling rate. We consider these as promising future works.

**Acknowledgement** The work was supported in part by the NSF IIS-1017362, NSF CNS-0931975, NASA NRA-NNH10ZDA001N, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, Boeing company, and by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References

Bar-David S, Bar-David I, Cross PC, Ryan SJ, Knechtel CU, Getz WM (2009) Methods for assessing movement path recursion with application to african buffalo in South Africa. Ecology 90:2467–2479

Berberidis C, Aref WG, Atallah MJ, Vlahavas IP, Elmagarmid AK (2002) Multiple and partial periodicity mining in time series databases. In: ECAI

Cao H, Cheung DW, Mamoulis N (2004) Discovering partial periodic patterns in discrete data sequences. In: PAKDD

Cao H, Mamoulis N, Cheung DW (2005) Mining frequent spatio-temporal sequential patterns. In: ICDM, pp 82–89

Cooke SJ, Hinch SG, Wikelski M, Andrews RD, Kuchel LJ, Wolcott TG, Butler PJ (2004) Biotelemetry: a mechanistic approach to ecology. Trends Ecol Evol 19:335–343

Cross PC, Lloyd-Smith JO, Johnson PLF, Getz WM (2005) Dueling timescales of host movement and disease recovery determine invasion of disease in structured populations. Ecol Lett 8:587–595

Dalziel BD, Morales JM, Fryxell JM (2008) Fitting probability distributions to animal movement trajectories: dynamic models linking distance, resources, and memory. Am Nat 172:248–0258

Elfeky MG, Aref WG, Elmagarmid AK (2005a) Periodicity detection in time series databases. IEEE Trans Knowl Data Eng 17(7):875–887

Elfeky MG, Aref WG, Elmagarmid AK (2005b) Warp: time warping for periodicity detection. In: ICDM, pp. 138–145

Getz WM, Saltz D (2008) A framework for generating and analyzing movement paths on ecological landscapes. Proc Nat Acad Sci USA 105(49):19066–19071

Han J, Dong G, Yin Y (1999) Efficient mining of partial periodic patterns in time series database. In: ICDE

Han J, Gong W, Yin Y (1998) Mining segment-wise periodic patterns in time-related databases. In: KDD, pp 214–218

Hewitson L, Dumont B, Gordon IJ (2005) Response of foraging sheep to variability in the spatial distribution of resources. Anim Behav 69:1069–1076

Indyk P, Koudas N, Muthukrishnan S (2000) Identifying representative trends in massive time series data sets using sketches. In: VLDB

Jensen CS, Lin D, Ooi BC (2004) Query and update efficient b+-tree based indexing of moving objects. VLDB, pp 768–779

Jeung H, Liu Q, Shen HT, Zhou X (2008) A hybrid prediction model for moving objects. In: ICDE

Lahiri M, Berger-Wolf TY (2008) Mining periodic behavior in dynamic social networks. In: ICDM, pp 373–382

Li Z, Ding B, Han J, Kays R (2010a) Swarm: mining relaxed temporal moving object clusters. Proc VLDB Endow 3(1):723–734

Li Z, Ding B, Han J, Kays R, Nye P (2010b) Mining periodic behaviors for moving objects. In: KDD, pp 1099–1108

Ma S, Hellerstein JL (2001) Mining partially periodic event patterns with unknown periods. In: ICDE
Mamoulis N, Cao H, Kollios G, Hadjieleftheriou M, Tao Y, Cheung DW (2004) Mining, indexing, and querying historical spatiotemporal data. In: KDD
McIntyre CL, Adams LG (1999) Reproductive characteristics of migratory golden eagles in Denali National Park, Alaska. In: JSTOR
McNaughton SJ (1985) Ecology of a grazing ecosystem: the Serengeti. Ecol Monogr 55:259–294
McNaughton SJ, Banyikwa FF, McNaughton MM (1997) Promotion of the cycling of diet-enhancing nutrients by african grazers. Science 278:1798–1800
Nathan R, Getz WM, Revilla E, Holyoak M, Kadmon R, Saltz D, Smouse PE (2008) Moving forward with movement ecology. Proc Nat Acad Sci USA 105(49):19052–19059
Patel JM, Chen Y, Chakka VP (2004) Stripes: an efficient index for predicted trajectories. In: SIGMOD conference, pp 637–646
Patternson TA, Thomas L, Wilcox C, Ovaskainen O, Matthiopoulos J (2008) State-space models of individual animal movement. Trends Ecol Evol 23(2):87–94
Polis GA, Anderson WB, Holt RD (1997) Toward an integration of landscape and food web ecology: the dynamics of spatially subsidized food webs. Ann Rev Ecol Syst 28:289–316
Saltenis S, Jensen CS, Leutenegger ST, Lopez MA (2000) Indexing the positions of continuously moving objects. In: SIGMOD conference, pp 331–342
Sugden A, Pennisi E (2006) When to go, where to stop: introduction. Science 313:775
Tao Y, Papadias D (2003) Spatial queries in dynamic environments. ACM Trans Database Syst 28(2):101–139
Tao Y, Papadias D, Sun J (2003) The tpr*-tree: an optimized spatio-temporal access method for predictive queries. In: VLDB 790–801
Tao Y, Faloutsos C, Papadias D, Liu B (2004) Prediction and indexing of moving objects with unknown motion patterns. In: SIGMOD conference, pp 611–622
Vlachos M, Yu PS, Castelli V (2005) On periodicity detection and structural periodic similarity. In: SDM
Wang C, Parthasarathy S (2006) Summarizing itemset patterns using probabilistic models. In: KDD
Wang W, Yang J, Yu PS (2001) Meta-patterns: revealing hidden periodic patterns. In ICDM
Wang Y, Lim E-P, Hwang S-Y (2003) On mining group patterns of mobile users. In: DEXA, pp 287–296
Wittemyer G, Polansky L, Douglas-Hamilton I, Getz WM (2008) Nonstationary influences of season, location and sociality on properties of movement among african elephants (*Loxodonta africana*). Proc Nat Aced Sci USA 105(49):19108–19113
Worton BJ (1989) Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70:164–168
Yan X, Cheng H, Han J, Xin D (2005) Summarizing itemset patterns: a profile-based approach. In: KDD
Yang J, Wang W, Yu PS (2000) Mining asynchronous periodic patterns in time series data. In: KDD
Yang J, Wang W, Yu PS (2002) Infominer+: mining partial periodic patterns with gap penalties. In: ICDM
Yang J, Wang W, Yu PS (2004) Mining surprising periodic patterns. Data Min Knowl Discov 9(2):189–216
Zhang M, Kao B, Wai-Lok CD, Yip KY (2005) Mining periodic patterns with gap requirement from sequences. In: SIGMOD conference