Taylor & Francis
Taylor & Francis Group

A new method for discovering behavior patterns among animal movements

Yuwei Wang[a,b], Ze Luo[b], John Takekawa[c,d], Diann Prosser[e], Yan Xiong[b],
Scott Newman[f], Xiangming Xiao[g], Nyambayar Batbayar[h], Kyle Spragens[c],
Sivananinthaperumal Balachandran[i] and Baoping Yan[b]

[a]University of Chinese Academy of Sciences, Beijing, China; [b]Computer Network Information Center,
Chinese Academy of Sciences, Beijing, China; [c]Western Ecological Research Center, US Geological Survey,
Vallejo, CA, USA; [d]Science Division, National Audubon Society, San Francisco, CA, USA; [e]Patuxent Wildlife
Research Center, US Geological Survey, Beltsville, MD, USA; [f]EMPRES Wildlife Health and Ecology Unit,
Animal Production and Health Division, Food and Agriculture Organization of the United Nations, Rome,
Italy; [g]Department of Microbiology and Plant Biology, Center for Spatial Analysis, University of Oklahoma,
Norman, Oklahoma, USA; [h]Wildlife Science and Conservation Center, Ulaanbaatar, Mongolia; [i]Bombay
Natural History Society, Hornbill House, Mumbai, India

ABSTRACT

Advanced satellite tracking technologies enable biologists to track animal movements at fine spatial and temporal scales. The resultant data present opportunities and challenges for understanding animal behavioral mechanisms. In this paper, we develop a new method to elucidate animal movement patterns from tracking data. Here, we propose the notion of continuous behavior patterns as a concise representation of popular migration routes and underlying sequential behaviors during migration. Each stage in the pattern is characterized in terms of space (i.e., the places traversed during movements) and time (i.e. the time spent in those places); that is, the behavioral state corresponding to a stage is inferred according to the spatiotemporal and sequential context. Hence, the pattern may be interpreted predictably. We develop a candidate generation and refinement framework to derive all continuous behavior patterns from raw trajectories. In the framework, we first define the representative spots to denote the underlying potential behavioral states that are extracted from individual trajectories according to the similarity of relaxed continuous locations in certain distinct time intervals. We determine the common behaviors of multiple individuals according to the spatiotemporal proximity of representative spots and apply a projection-based extension approach to generate candidate sequential behavior sequences as candidate patterns. Finally, the candidate generation procedure is combined with a refinement procedure to derive continuous behavior patterns. We apply an ordered processing strategy to accelerate candidate refinement. The proposed patterns and discovery framework are evaluated through conceptual experiments on both real GPS-tracking and large synthetic datasets.

CONTACT Ze Luo ✉ luoze@cnic.cn

1. Introduction

The technical advancement of positioning technologies and the increasing pervasiveness of location-acquisition devices have led to large quantities of movement data from various types of moving objects, such as humans, animals, vehicles, and natural phenomena. The trajectory data collected from these objects provide researchers with an opportunity to obtain valuable knowledge about individual or collective behaviors. This knowledge can be of practical use in a variety of applications, such as social network, traffic control, mobile recommendation, and biological study. In recent ecological research studies, satellite tracking technologies have played a significant role in the investigation of animal movement dynamics in large spatial and temporal scales; the use of these technologies has become a popular method for remotely clarifying animal behaviors (Bridge et al. 2011, Cui et al. 2011).

Annual movement patterns can be described as a series of movements among discrete locations on the landscape that are shaped by the resource availability required at a distinct time stage in the annual life cycle of species (Alerstam and Lindström 1990, Newton 2008). These movements are observed at certain spatial and temporal scales depending on species (Fretwell and Lucas 1969, Johnson 1980, Drent et al. 2003). For example, arctic breeding goose populations migrate over long distances to temperate latitudes to take advantage of food resources (Owen 1980, Prop et al. 2003). Individuals exhibit tendencies to return to sites visited previously (Black et al. 2007); however, the location of these sites varies within a species at the population level (Takekawa et al. 2009). Animal movement patterns can mainly be derived according to the spatiotemporal information provided by the locations visited by the animal. Migration tracking data provide crucial insights that answer basic questions in biological studies and that explain observations of animal migration and life histories. These insights have long eluded scientists (Bridge et al. 2011). These data can also assist in identifying animal residency (Pedersen et al. 2011), the connectivity between breeding and wintering areas, and further guide conservation efforts for preserving animal populations within regional landscapes (Faaborg et al. 2010, Bridge et al. 2011). However, insights into animal behavior must be refined as a new challenge in data management and analyses that utilize high-frequency and high-accuracy datasets.

Many data mining technologies can be applied to animal movement data to abstract animal mobility-related phenomena, such as in studies on spatial co-locations (Zhang et al. 2004) and group relationship patterns (Benkert et al. 2008, Jeung et al. 2008, Tang et al. 2013), spatiotemporal sequential patterns (Cao et al. 2005, Giannotti et al. 2007), and periodic patterns (Li et al. 2011). These contributions could serve as the basis of further studies on the intrinsic link between animal movement and the long-term spread of disease (Prosser et al. 2009, Takekawa et al. 2010, Altizer et al. 2011), and the anticipated responses to changes in environmental conditions and climate. In turn, behavioral states and even associated locations can also be estimated from the environmental and other factors (Pedersen et al. 2011).

In this article, we propose an expanded version of the spatiotemporal sequential pattern to directly represent the underlying collective behavioral states and the sequential modes of migratory animals, called Continuous Behavior Pattern (CB-Pattern). The CB-Pattern is derived from multiple migration trajectories that are similar in the space

and time dimensions throughout the life span of the pattern. Moreover, the CB-Pattern is a spatiotemporal representation of popular migration behavior sequences shared by multiple individuals. Each element in such a pattern is composed of the common underlying behaviors in multiple individuals during migrations and corresponds to a specific life phase. These behaviors are characterized by representative spots with space and stay duration properties. The behavioral state (e.g., breeding, wintering, molting, and stopover) of each element can be inferred from the spatiotemporal description of individual behaviors. Hence, the CB-Pattern can help biologists identify similar individuals and popular migration routes, understand animal behavioral modes and the phases of migration life, and guide regional conservation efforts directed at preserving animal populations during specific phases of annual animal requirements.

To discover the CB-Patterns from raw migration movement data (trajectory data), we develop a Candidate Generation and Refinement (CGAR) framework. We first define the representative spot as a relaxed continuous segment in the trajectory. This segment corresponds to a single migration behavior. We then extract the sequence of representative spots with space and stay duration properties from the trajectory. Similar migration behaviors are often exhibited in almost-identical locations and at like durations. Given the spatiotemporal proximity of representative spots, we utilize clustering technology (Ester *et al*. 1996) to identify similar representative spots, which reveal specific and significant life states during migration. We then connect the relaxed continuous clusters of spots to form candidate CB-Patterns based on a traditional sequential pattern mining paradigm according to the sequential relationships among the spots in time dimension. When the continuity constraint is relaxed, the CB-Pattern contains closely linked sequential behaviors, and tolerates slight divergences or disturbances in real-life, homogeneous individual movements. The candidates are then refined by removing the spots from the movements that are inconsistent with the patterns in certain phases, and by updating the clustering structure at each phase of the pattern. Moreover, we propose a simple but efficient processing strategy to accelerate the refinement procedure of this framework.

Extensive simulated experiments are conducted on real GPS location and large synthetic datasets to validate our proposed approaches as well as to facilitate a comprehensive analysis of the experimental results. This analysis demonstrates the interpretability of the patterns discovered (Wang *et al*. 2014). In this article, we simplify the definition of the cluster, thereby redefining the proposed animal behavior pattern, and further improving on the framework for pattern discovery. We also conduct a new analysis on the ecological values of the pattern discovery problem and on the experiments performed within this new framework.

2. Problem definition

In this section, we state the related concepts. The main notations used throughout this article are listed in Table 1.

Definition 2.1:   A **trajectory** T is a series of points $\langle p_1, p_2, \ldots, p_n \rangle$ arranged in ascending order of time. These points are collected from an animal and each point possesses spatial coordinate and time stamp properties.

Table 1. List of notations used in this article.

Notation	Explanation	Notation	Explanation
T, T_i, T_j	Trajectories	$sp.d$, $sp.c$	Duration and center of sp
T_{DB}	Trajectory dataset	EPS	Spatiotemporal radius
p, p_i, p_j	Points in trajectories	$MinPts$	Density threshold in clustering
sp, sp_i, sp_j	Representative spots	$Ner_D(sp)$	Neighborhood of sp in the spot set D
S, S_i, S_j	Spot sequences	$MaxI$	Time interval threshold in the CB-Pattern
S_{DB}	Set of spot sequences	δ_G, δ_D	Gap and distance thresholds for spots
C, C_i, C_j	The spot clusters	$MinL$	Minimum length of a CB-Pattern
P, P_i, P_j	CB-Patterns	$dist(p_i, p_j)$	Spatial distance between p_i and p_j
$P.L$	Length of P	$S|_P$	Projection of S with respect to P
$sp.s$, $sp.e$	Starting time and ending time of sp	$S_{DB}|_P$	Projection dataset with respect to P

In this article, behavior refers to a specific behavior or annual cycle phase, such as breeding, molting, stopover, and flyover. To identify behavior-related patterns, we must consider how potential migration behaviors can be derived from raw trajectories. In addition to the geographic locations visited by animals, the temporal context is important for understanding underlying behavioral states. For example, a region can be used by various migratory birds or different populations of the same species (e.g., bar-headed goose, *Anser indicus*, Takekawa *et al*. 2009) as a breeding, molting, or stopover site, as with the Qinghai Lake region in China. The residence of bar-headed geese within this region for approximately 1 month after the spring migration indicates potential breeding behavior; an extended stay until the autumn migration can reflect a molting phase; a stay for only several days may correspond to a stopover event during which these birds rest and refuel for continued migration (Takekawa *et al*. 2009, Cui *et al*. 2011). Hence, important cues to understanding and defining recorded underlying behaviors are the location in which behaviors are observed, the duration of these behaviors, and the timing related to the annual life history strategies of species. We apply the representative spot notion to characterize potential behavior in terms of location and duration. The spot is defined as a movement segment in a trajectory as follows.

Definition 2.2: Given trajectory $T = \langle p_1, p_2, \ldots, p_n \rangle$, distance threshold δ_D, and gap threshold δ_G, a *representative spot* (hereafter *spot*) sp consists of maximal (possibly non-consecutive) subsequence $\langle p_{k_1}, p_{k_2}, \ldots, p_{k_m} \rangle$ of T, where $0 < k_i - k_{i-1} \leq \delta_G$ ($1 < i \leq m$) and $dist(p_{k_i}, sp.c) \leq \delta_D$ ($1 \leq i \leq m$). The starting and ending times are defined as the time stamps of the first and the final points in the subsequence, and the interval length of the subsequence reflects spot duration.

The spot suggests that an animal remains in a spatial extent for a period and indicates potential significant behavior for a specific life need. The concept of the spot is similar to that of potential stop as proposed by Trasarti *et al*. (2011) and that of stay point as presented by Li *et al*. (2008). Nonetheless, two important differences are observed. First, the points in a potential stop or a stay point must be strictly continuous. Noise in the trajectories can be attributed to the occasional failures of GPS devices and other interferences. In this case, the previously developed approaches split a complete behavior into several segments. Second, these methods consider only the places at which an animal stays for a sufficiently long period of time, and ignore short-interval migration phases such as stopovers, which are also important in investigating animal migration dynamics. We consider the distance to the center instead of the first point in

the behavior because animals often display a pendulum-like movement relative to a nest site during breeding season.

According to Definition 2.2, a spot may be generated from individual random movements. If multiple individuals generate spots that are similar in location and time dimensions, these spots should indicate a special type of significant common behavioral state. Hence, we apply clustering technology to identify these like spots and capture the common behaviors shared by multiple animals. To determine the cluster exhibiting a spatiotemporal extent of arbitrary shape, we extend the notion of density-based clustering (Ester *et al.* 1996) to the spatiotemporal context (called *ST-DBSCAN*) by employing the following spatiotemporal neighborhood definition. Let *EPS* be a spatiotemporal radius with spatial component *EPS.d* and time component *EPS.t* while *MinPts* is a density threshold. The neighborhood of spot *sp* is defined as the set of spots whose spatial distance to *sp* is no greater than *EPS.d*, and the difference in terms of the duration is no more significant than *EPS.t*. A spot is a core if the size of its neighborhood reaches *MinPts*. Spot sp_j is directly density-reachable from sp_i if *O* is a core and sp_j is in the neighborhood of sp_i. Spot sp_j is density-reachable from sp_i, if there is a chain of spots $sp_{k_1}, sp_{k_2}, \ldots, sp_{k_m}$ ($sp_j = sp_{k_1}, sp_i = sp_{k_m}$) such that sp_{k_l} is directly density-reachable from $sp_{k_{l+1}}$ ($\forall 1 \leq l < m$). Furthermore, sp_i and sp_j are density-connected to each other, if there is sp_k such that both sp_i and sp_j are density-reachable from sp_k. Finally, a *spot cluster* (hereafter cluster) *C* is a non-empty set of spots such that any two sp_i, sp_j in *C* are density-connected.

A cluster is a rough spatiotemporal outline of certain common behaviors. In the initial clustering stage, the clusters identified on all spots may contain many occasional spots that originate from the random movements of many individuals and do not actually exhibit underlying unified behaviors. However, if multiple individuals display a similar spot sequence over a long period, then these individuals display categorically similar behavioral states in each phase of this period. In addition, certain closely linked sequential behavioral states are repeated over the course of the life span of a particular species or population (Pedersen *et al.* 2011). For example, a breeding state is often closely followed by a molting state in waterfowl. Hence, we aim to identify the long-term sequential multiple behaviors that are observed occur in the movements of multiple individuals. The CB-Pattern contains a sequence of clusters to represent this type of reoccurring behavior.

Definition 2.3: Let *MaxI* be the time interval and *T* be a trajectory from an individual. A spot sequence is a sequence of spots from *T*, which is denoted as $S = \langle sp_1, sp_2, \ldots, sp_n \rangle$. In this equation, $sp_i.e < sp_{i+1}.s$. If $sp_j.s - sp_i.e \leq MaxI$ for the sp_i and sp_j in *S*, then sp_j is a linked successor of sp_i. Let $S' = \langle sp_{k_1}, sp_{k_2}, \ldots, sp_{k_m} \rangle$ be a (possibly non-consecutive) subsequence of *S*. S' is a *continuous spot sequence* if $sp_{k_{l+1}}$ is a linked successor of sp_{k_l} for $1 \leq l < m$.

Definition 2.4: A CB-Pattern with length *L* ($L \geq MinL$ and *MinL* is the pattern length threshold) is denoted by a sequence of spot clusters $P = \langle C_1, C_2, \ldots, C_L \rangle$, which are formed by a set of spot sequences (called *instances* of *P*). The CB-Pattern satisfies the following criteria: (1) each instance is a continuous spot sequence with *L* spots under the

threshold *MaxI*; (2) each C_l is the set of the *l*-th spots of all instances, and the set is a spot cluster with respect to parameters *EPS* and *MinPts*; (3) if there is a continuous spot sequence $\langle sp_{k_1}, sp_{k_2}, \ldots, sp_{k_L} \rangle$ such that $C_l \cup sp_{k_l}$ $(\forall 1 \leq l \leq L)$ is a cluster, then $sp_{k_l} \in C_l$ and $\langle sp_{k_1}, sp_{k_2}, \ldots, sp_{k_L} \rangle$ is an instance of P (i.e., P contains as many instances as possible).

The spots in cluster C_l of the CB-Pattern P should be density-connected to one another without the aid of the spots beyond C_l. Thus, the initial clusters on the set are all original spots that can include many occasional spots from animals whose movements accidentally intersect with the pattern. In the CB-Pattern, we eliminate the interference of these occasional spots, and aim to derive accurate spatiotemporal descriptions of the underlying behavior sequence as well as individual movements that comply with the entire pattern. In Definition 2.4, the CB-Pattern can tolerate frequent slight divergences or disturbances in real-life similar movements because of the application of relaxed continuity constraint *MaxI*. In the following paragraph, we present the projection-related concept that is extended from the projection-based sequential pattern mining algorithm PrefixSpan (Pei *et al*. 2001) and applied to the pattern discovery framework.

Definition 2.5: Given animal spot sequence $S = \langle sp_1, sp_2, \ldots, sp_n \rangle$ and CB-Pattern P, we assume that the subsequence $S' = \langle sp_{k_1}, sp_{k_2}, \ldots, sp_{k_L} \rangle$ in S is an instance of P and that the last spot sp_{k_L} in S' is the *i*-th spot sp_i in. The subsequent sequence in S, i.e., $\langle sp_{i+1}, sp_{i+2}, \ldots, sp_n \rangle$, is a *projection* of S with respect to P. If there is no instance of P in S, then the projection of S with respect to P is empty. If multiple instance are detected in S, then multiple projections of S are generated. The projection set of S with respect to P is written as $S|_P$. The projection set of all animal spot sequences with respect to P is expressed as $S_{DB}|_P$.

We provide a simple example of the projection as follows. Assume that $MaxI = 3$, the current CB-Pattern $P = \langle a, b \rangle$ (a and b are the cluster identifiers), and a spot sequence S is $\langle (a, 1, 3), (b, 5, 8), (c, 9, 10), (d, 11, 12), (a, 14, 17), (e, 18, 18), (b, 19, 21), (b, 23, 24) \rangle$. In this example, a triple indicates a spot which contains the cluster identifier, the starting time and ending time of the spot. Then, $S|_P = \{ \langle (c, 9, 10), (d, 11, 12), (a, 14, 17), (e, 18, 18), (b, 19, 21), (b, 23, 24) \rangle, \langle (b, 23, 24) \rangle \}$.

3. Pattern discovery framework

In this section, we present a *CGAR* framework for discovering all CB-Patterns, and propose an ordered processing strategy to accelerate this framework (CGAR using the Ordered processing strategy OS is called *CGAR_OS*). The CB-Pattern is built on the concept of representative spots. Therefore, we first present the algorithm to extract the spot sequence (Algorithm 1). According to Definition 2.2, two or more spots from the same trajectory may coexist and overlap in the time dimension. To avoid this situation, we adopt an early selection principle (Palma *et al*. 2008). Algorithm 1 scans the raw trajectory once and returns a spot sequence in which each spot has the following attributes: center, starting time, ending time, and duration time. The settings of parameters δ_G and δ_D depend on the specific application scenario. For example, when we consider the movement of migratory birds, δ_D may be set to several tens of kilometers and δ_G to a small integer.

Algorithm 1: Spot extraction

Input: A trajectory T, δ_G, δ_D

Output: A spot sequence S

1:    $i = 0$, $S = <>$;
2: **while** $i < length(T)$ **do**
3:      $j = i + 1$, $skipped = 0$, $lastIndex = i$;
4:      $sp = \langle p_i \rangle$; // start a new spot
5:      **while** $j < length(T)$ **do**
6:        **if** $skipped \geq \delta_G$ **then** // end of the current spot
7:          $setStartEndTime(sp)$;
8:          $S.Append(sp)$;
9:          **break**;
10:       **else if** $dist(p, sp.c) \leq \delta_D$ **then**
11:         $sp.Append(p_j)$;
12:         $lastIndex = lastIndex + 1$; // reset the number of point skipped continuously
13:       **else** // the current point is consider as the noise
14:         $skipped = skipped + 1$;
15:       **end if**
16:         $j = j + 1$;
17:      **end while**
18:      $i = lastIndex + 1$;
19: **end while**
20: **return** S

The CGAR framework is presented in Algorithm 2. CGAR consists of the spot extraction phase and two intertwined phases, namely, candidate generation and candidate refinement. The candidate generation phase (Lines 6 ~ 15) follows the pattern-growth strategy in the PrefixSpan algorithm (Pei *et al*. 2001), where candidate CB-Patterns (i.e., a sequence of clusters) are generated incrementally. We extend the current CB-Pattern P to a long candidate P_{cand} using the clusters identified on the linked successor spots of the final spots in the current instances of P. That is, we extend the current instances with linked successor spots. In Line 15, some current instances cannot be extended and are thus inconsistent with P_{cand}; consequently, these spots are removed. Upon eliminating the interference of these occasional spots, the clustering structures of the changed clusters in P_{cand} may be altered. One initial cluster in P may evolve into either an accurate cluster or several separable clusters. Hence, we refine the candidates to derive true CB-Patterns in the candidate refinement phase (Line 16).

Algorithm 2: The Candidate Generation and Refinement Framework (CGAR)

Input: A set of trajectory T_{DB}, δ_G δ_D,, EPS, $MinPts$, $MaxI$ and $MinL$

Output: A set of CB-Patterns R

1: $S_{DB} = \emptyset$; // initialize a set of spot sequences
2: **for** each $T \in T_{DB}$ **do**

3:    $S = extractSpots(T, \delta_G, \delta_D)$; //extract the spot sequence

4:    $S_{DB} = S_{DB} \cup S$;

5: **end for**

6: $Q = \{(<>, S_{DB})\}$; // initialize a candidate $P = \langle \rangle$ with length 0 and S_{DB} as its projection dataset, and use a queue Q to store the pairs

7: **while** $Q \neq \emptyset$ **do**

8:    $(P, S_{DB}|_P) = Q.DeQueue()$;

9:    **if** $P.L \geq MinL$ **then**

10:       $R = R \cup P$;

11:    **end if**

12:    $SP_{next} = collectSuccessors (S_{DB}|_P, MaxI)$;

13:    $CS = ST\_DBSCAN(SP_{next}, EPS, MinPts)$; // cluster the spots using the modified DBSCAN algorithm

14:    **for** each $C \in CS$ **do**

15:       $P_{cand} = extendInstances(P, C)$.; // generate a candidate by appending the spots in to the instances of P

16:       $PS = refine(P_{cand}, EPS, MinPts)$; // PS is the set of true CB-Patterns derived from P_{cand}

17:       **for** each $P_{next} \in PS$ **do**

18:          $S_{DB}|_{P_{next}} = projecting(P_{next})$;

19:          $Q.EnQueue((P_{next}, S_{DB}|_{P_{next}}))$; // add the new pair to Q

20:       **end for**

21:    **end for**

22: **return** R;

In the candidate generation phase of Algorithm 2, the subsequent clusters for extending P are identified on the projection dataset with respect to P. Here, we employ the pseudo-projection technique (Pei *et al.* 2001) to reduce the cost of projection; that is, we use the pointer referring to the spot sequence and the offset the final spot of the pattern instance in the sequence instead of storing the true projection subsequence. In the function *collectSuccessors* (Line 12 in Algorithm 2), we obtain all linked successors of the last spots in the instances of P. Then, we cluster these spots using the ST_DBSCAN clustering algorithm mentioned in Section 2. We append the spots in each cluster to the instances of P, and generate a new candidate P_{cand} in the process (Line 13 in Algorithm 1).

If certain spots are removed from the *l*-th ($1 \leq l \leq P.L$) cluster when P is extended to P_{cand}, then the refinement procedure is conducted (Line 14 in Algorithm 1). The steps of refinement procedure are shown in Algorithm 3, which updates candidate P_{cand} and may divide it into several CB-Patterns. If the *l*-th spot set C_l changes after the final clustering process performed on it, we examine whether or not the remaining spots in C_l still constitute a cluster. If spot set C_l is divided into several clusters, then a new candidate is generated with respect to each newly generated cluster (Line 8 in Algorithm 3), and the refinement procedure is recursively executed on each new candidate(Line 9 in Algorithm 3). Each candidate either evolves into a CB-Pattern or is abandoned if any of the sets no longer contains a cluster.

Algorithm 3: Refine

Input: A candidate CB-Pattern.P_{cand}., EPS, MinPts

Output: A set of CB-Patterns

1: **for** $C_i \in P_{cand}$ **do**

2:     **if** isChanged(C_i) = false **then** // C_i is unchanged

3:         continue;

4:     **end if**

5:     $CS = ST\_DBSCAN(C_i, EPS, MinPts)$;

6:     **if** $CS \neq C_i$ **then** // the clustering structure on C_i has changed

7:         **for** $C \in CS$ **do**

8:           $P'_{cand} = SplitWith(P_{cand},\ C)$; // Retain the instances associated with the spot in C

9:           $PS' = Refine(P_{cand},\ EPS,\ MinSup)$;

10:           $PS = PS \cup PS'$;

11:         **end for**

12:     **end if**

13:     **return** PS;

14:**end for**

15:**if** $i > P_{cand}.L$ **then** // each spot set in P_{cand} is a cluster

16:     **return** $\{P_{cand}\}$;

17:**end if**

The clustering algorithm ST_DBSCAN is used frequently in the CGAR framework. During the clustering process, for a spot we must scan the spot set to retrieve its neighborhood. It takes $O(n^2)$ time for n spots without a spatial index. Maintaining a spatial index, such as R-tree, leads to a high cost of processing time because the set of spots always changes (Tang *et al.* 2013). This, we employ a simple three-dimensional (i.e., latitude, longitude and time) grid structure as the index by partitioning time and space into cubes with widths EPS.d and EPS.t in the space and time dimensions, respectively. We then directly retrieve the spots located in adjacent cells and filter these spots as the neighborhood during the neighborhood search for a spot. The grid can be maintained in $O(n)$ time.

During the refinement procedure, we examine all spot sets sequentially, as shown in Algorithm 3. We can accelerate this procedure through heuristic strategies. In the refinement phase, we can determine the change in each spot set since the previous clustering on this set, i.e., we can detect the numbers removed from this set. The number of spots removed is often unequal to the number of instances eliminated, and the numbers may be inconsistent for different sets of one candidate pattern. On the basis of the number of spots removed from each set, we design an ordered processing strategy to accelerate candidate refinement. In the employed strategy, the cluster structures of the sets are examined in the descending order of the numbers of elements removed from the sets. A significant change in a set may lead to greater shrinkage of the cluster, and therefore, more interrelated spots in the candidate are removed. Hence, the candidate can be updated more quickly and evolves into one or more true patterns, or is abandoned once any of its sets does not contain a cluster. The CGAR framework that use the Ordered processing Strategy is called *CGAR_OS*, as mentioned previously.

4. Experiments

In this section, we present the results of our proposed CB-Pattern given real trajectory datasets of wild birds. Specifically, we design a trajectory generator to generate large synthetic datasets and to systematically evaluate the running cost of the framework. All the algorithms used in these experiments are implemented in Java and are executed on a single machine with Intel Core(TM) 2 Quad 2.66 GHz CPU and 4 GB memory.

4.1. Datasets

We obtained three real trajectory datasets from three separate satellite tracking projects comprised of multiple deployments for wild birds marked in China, India, and Mongolia. The tracking projects in China were deployed in 2007 and 2008. They generated 55 trajectories from 29 bar-headed geese and 26 ruddy shelducks. The tracking projects in India were deployed from 2005 to 2011, and resulted in 36 trajectories of bar-headed geese and 14 trajectories of ruddy shelducks. The tracking project in Mongolia was deployed in 2008 and resulted in 38 bar-headed geese. In summary, the aggregate datasets contained 143 trajectories of wild birds for use in this analysis. The two monitored species, bar-headed goose and ruddy shelduck, migrate along the migration corridors in Central Asia. These birds were equipped with solar-powered, GPS-accuracy Platform Terminal Transmitters. The locations were estimated either through GPS or with the Argos system when a GPS-fix cannot be obtained. These datasets are available for visualization on the website of U.S. Geological Survey.[1] We sampled the location data at a frequency of one per day. The available length of a generated trajectory for an individual bird varies from several days to 3 years, and the total number of location records is 31,324. An overview of the dataset is shown in Figure 1(a).

In addition to the real dataset, we designed a trajectory generator that synthesizes new trajectories based on the real trajectory dataset. Data generation is controlled by the following parameters: number N of trajectories, average length L of trajectories, ratio
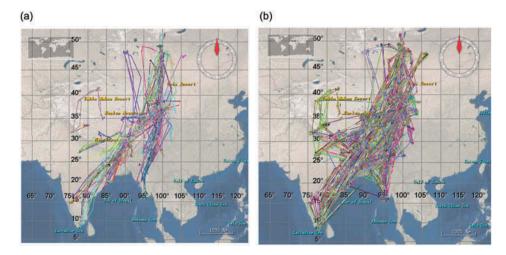

Figure 1. Overview of the trajectory datasets. Each colored polyline indicates a trajectory. (a) Real trajectory datasets of wild birds. (b) Example of synthetic datasets generated with the developed trajectory generator.

$α$ of noisy trajectories, spatial standard deviation $γ$ and temporal standard deviation $ε$ to add perturbations to the reference spot, standard deviation $σ$ of random movements relative to the spot, the possibility $β$ that an object changes its reference trajectory during movement, and agility $λ$ of the movement among consecutive reference spots. In the generation process, we applied the real spot sequences of the birds as references. To generate a trajectory, we randomly select the spot sequence of an animal as the reference sequence and utilize its initial spot as the reference spot. We introduced the variance by perturbing the spot through bivariate Gaussian distribution with standard deviations $γ$ and $ε$. In addition, we further randomly generate trajectory points relative to the perturbed spot via Gaussian distribution with standard deviation $α$. Then, the object moves to the next spot along the reference sequence. During this iteration, we generate in-between locations in a random direction under the control of agility threshold $λ$. The agility threshold $λ$ denotes the maximum deviation between actual movement direction and the linear direction from the current location to the destination spot. The reference trajectory of an object can change during the generation process with possibility $β$, and alters changes its movement patterns thereafter. During the change, we randomly select a spot associated with other animals according to the bivariate Gaussian distribution on the current spot with spatial standard deviation $γ$ and temporal standard deviation $ε$. The associated sequence is regarded as new reference and the next spot as the new reference spot. The dataset contains $α × N$ random trajectories which are randomly generated with arbitrary spots. The default parameters of the data generator used in these experiments are as follows: $N = 2000$, $L = 800$, $α = 0.1$, $.γ = 25$. km, $ε = 5$ days, $σ = 15$ km, $β = 0.1$, and $λ = 45°$. Hence, the default size of the synthetic GPS dataset is 1,600,000. An example of the synthetic dataset is also presented in Figure 1(b).

4.2. Results

We first present the experimental results for the three real datasets. All the birds in the datasets are located in the migration corridors of Central Asia from Mongolia to India and their trajectories intersect partially. Thus, we apply the CGAR framework to these datasets simultaneously and report the corresponding results. On the basis of the related ecological study on the home ranges of these wild birds (Cui et al. 2011), we set $δ_D$ as 30 km and $δ_G$ as 3. Moreover, we set MinPts as 3, MaxI as 15 days, MinL as 1, EPS.d as 80 km, and EPS.t as 15 days.

We present five typical CB-Patterns derived from the bird dataset in Figure 2. These patterns are consistent with those highlighted in ecological research studies conducted in this region (Takekawa et al. 2009, Prosser et al. 2011), and cover the main migration routes of bar-headed geese and ruddy shelduck in Central Asia. These underlying behavioral states are inferred according to the location semantics and the time periods in which the clusters exist, as well as the close sequential relationship among the clusters in a pattern. For example, the longest pattern $P_1 = 1, 2, 3, 4, 5, 6, 7, 8$ of length 8 with the duration sequence <51 ~ 60, 1 ~ 3, 2 ~ 23, 1 ~ 13, 7 ~ 20, 1 ~ 13, 1 ~ 8, 1 ~ 9> (the green path in Figure 2) indicates that the corresponding birds breed in Region 1 for 51 ~ 60 days (Terkhiin Tsagaan Lake, Mongolia, 48.1629°N,99.6909°E), and then stop-over in Regions 2 ~ 8 (Khangai Nuruu National Park, northeast of Hala Lake, east of Hala
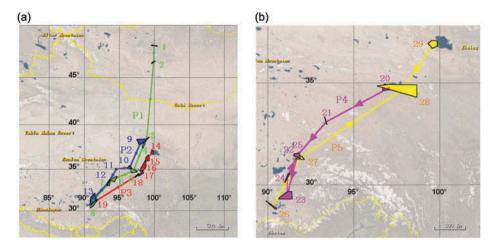
Figure 2. Examples of CB-Patterns derived from the wild bird satellite telemetry datasets. Each cluster in a CB-Pattern is represented by a polygon (convex hull is used here), and its identifier is marked near the polygon. Each directed path that connects the clusters denotes a CB-Pattern. Each path and the associated clusters are marked in a distinctive color. (a) Three CB-Patterns and (b) two CB-Patterns.

Lake, Zhaling Lake, Zhiduo County, Zhamucuo Wetland, and Selincuo BNC NR) successively. Pattern P_2 (the blue one in Figure 2) denotes a successive stopover sequence from Hala Lake (38.65°N,98.09°E), Qinghai Province, to Selincuo BNC NR (31.86°N,88.98°E). Pattern P_3 = 14, 15, 16, 17, 18, 19 with the duration sequence 145 ~ 170, 1 ~ 6, 1 ~ 4, 25 ~ 46, 1 ~ 10, 1 ~ 12 suggests some birds breed and molt in the Qinghai Lake region, and then continue to migrate south continuously. Pattern P_4 = 20, 21, 22, 23, 24, 25 with the duration sequence 20 ~ 42, 1 ~ 1, 6 ~ 19, 126 ~ 138, 1 ~ 2, 6 ~ 19 in Figure 2 shows that a flock of birds remains at Zhaling Lake before the start of autumn migration, passes Zhamucuo Wetland and Selincuo BNC NR, winters in Tibet River Valley (Region 16) for more than 126 days, and then migrates back to Namucuo Lake and then to Zhamucuo Wetland. Pattern P_5 = 26, 27, 28, 29 with the duration sequence 1 ~ 27, 1 ~ 12, 1 ~ 7, 116 ~ 136 in Figure 2 denotes a northward migration from Shannan Prefecture of Tibet Autonomous Region (29.26°N,91.76°E) in China during spring, before settling at the Qinghai Lake region in the breeding and molting periods.

By observing the individuals involved in these CB-Patterns, we detect significant differences that between the birds marked in Mongolia and those marked in China in terms of the selection of breeding sites. We also noted a slight variation between these two types of birds with regard to the migration routes taken. The birds marked in China bred at Qinghai Lake, China, whereas the majority of those marked in Mongolia fly over Hala Lake (northwest of Qinghai Lake) to Mongolia to breed. The birds marked in India are similar to those marked in Mongolia with regard to their behavior sequences. A few birds marked in India intersect with the birds marked in China: the spatial and temporal overlaps are concentrated in the migration periods. Consider the examples of CB-Patterns in Figure 2. P_1 is formed by the birds marked with GPS devices in Mongolia, whereas P_3 and P_4 are generated by the birds marked in Qinghai Lake, China. This phenomenon reveals the high discriminating power of CB-Patterns on different
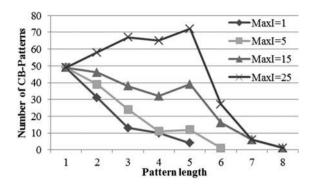
Figure 3. Number and length of CB-Patterns derived from the real datasets with respect to *MaxI*.

migration dynamics. P_5 is attributed to the birds marked in India and in Qinghai Lake, China, whereas P_2 involves the birds marked in Mongolia and India. This finding suggests the similarities and intersections among the birds marked in India and those from Mongolia and China.

Figure 3 shows the numbers of CB-Patterns derived from the wild bird datasets with respect to time interval threshold *MaxI* between two consecutive behaviors in these patterns. When *MaxI* is set to 1, no gap is allowed in the pattern. Many interesting patterns such as $P_1{\sim}P_5$ in Figure 2 can be identified by implementing the loose continuity constraint. This constraint facilitates the robustness of the CB-Pattern in the event of the missing or noisy data and the tolerance of the pattern for slight divergences in actually similar individual behaviors, and in the tracking data. Figure 4 presents the maximum, minimum, and average life spans of the patterns with respect to the pattern length. Here, we define pattern life span as the duration time from the start of a pattern to its end. Many significant, long-term similarities are observed in the movements of these birds.

4.3. Cost analysis

On the basis of the synthetic datasets, we studied the running cost of our proposed CGAR framework on CB-Pattern discovery. In particular, we measure the runtime cost and memory requirement of CGAR given different parameter settings and assess the
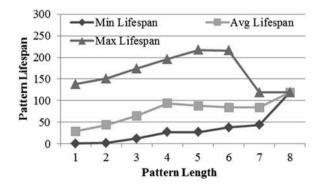

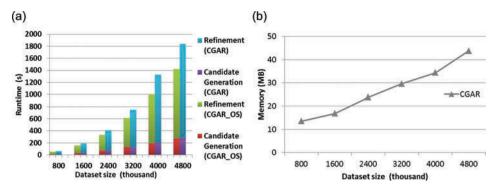**Figure 4.** Life spans of the patterns with respect to the pattern length.

Figure 5. Algorithm performance in relation to the size of the GPS dataset. (a) Runtime. (b) Memory requirement.

performance of both CGAR and CGAR_OS. Based on the observations on the experimental data, we set the parameters as: $\delta_D = 30$ km, $\delta_G = 3$, $MinPts = 0.02 \times N$, $MaxI = 11$ days, $MinL = 2$, $EPS.d = 60$ km, and $EPS.t = 8$ days.

As shown in Figure 5(a), we assess the runtimes of CGAR and CGAR_OS with respect to the size of synthetic GPS dataset by tuning the number of trajectories N in the trajectory generator. We calculate the times consumed by the candidate generation process and the refinement process separately. Runtime is sensitive to the dataset size, as expected. Moreover, the time costs of CGAR and CGAR_OS increase almost quadratically. This finding is attributed to the fact that the main cost of pattern discovery is associated with the clustering steps and that the time complexity of the density-based clustering algorithm used in this study is $O(n^2)$ for n spots. The runtime of the refinement process dominates the overall runtime of the framework. In general, the composition of the runtime is approximately 80% for refinement and roughly 20% for candidate generation because the refinement must be performed multiple times after a candidate is generated. As indicated in Figure 6(a), CGAR_OS outperforms CGAR by 18%~26% because of the selective processing order of the former in the refinement phase. In Figure 5(b), we show only the memory requirement of CGAR because the memory allocation is consistent between CGAR and CGAR_OS. The memory requirement of CGAR is low and increases linearly with the size of the input dataset.

We investigate the performance levels of CGAR and CGAR_OS with respect to *MinPts*, *MaxI*, and *EPS*. The results are depicted in Figures 6 and 7. As expected, CGAR_OS outperforms CGAR significantly in terms of runtime cost when these parameters are varied. As *EPS* and *MaxI* increase or *MinPts* decreases, the runtime of CGAR increases more quickly than that of CGAR_OS. The changes in the runtime costs of both CGAR and CGAR_OS are linear with respect to *MinPts* and *MaxI*. The rate of this increase slows down when *MaxI* increases from 6 days to 11 days. This outcome indicates that once *MaxI* has attained a certain value (6 ~ 11 days in this study), slight divergences in similar movements can be tolerated. Moreover, the runtimes of both CGAR and CGAR_OS increase significantly when *EPS.d* exceeds 60 km or when *EPS.t* exceeds 8 days. This result can be explained by the three-sigma rule in a Gaussian distribution (Pukelsheim 1994). In the data generator, the spatial standard deviation γ and the temporal standard deviation ε of the Gaussian distribution on the spots are set to 25 km and 5 days,
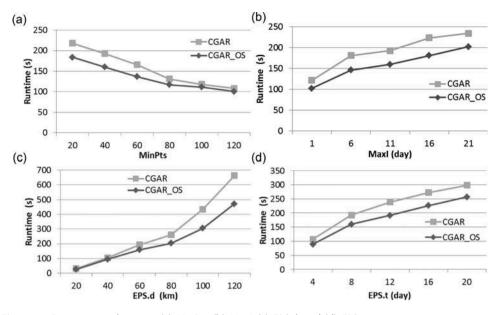
Figure 6. Runtime in relation to (a) *MinPts*, (b) *MaxI*, (c) *EPS.d*, and (d) *EPS.t*.
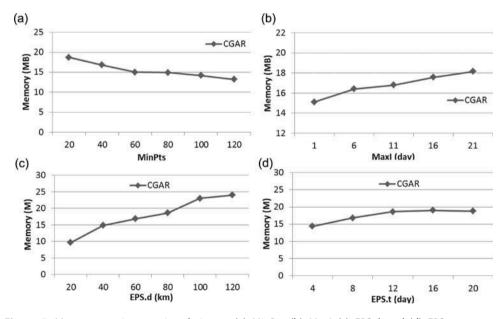

Figure 7. Memory requirement in relation to (a) *MinPts*, (b) *MaxI*, (c) *EPS.d*, and (d) *EPS.t*.

respectively. When the distance threshold exceeds several times the standard deviation, many of the elements mixed into the current cluster that are actually unrelated to this cluster. As a result, the clustering cost increases rapidly. As per Figure 8, the memory requirement is always low under different settings, and the change trend of this requirement is similar to that of runtime cost.

Figure 8. Number of CB-Patterns in relation to *MinPts*, *MaxI*, *EPS.d*, and *EPS.t*.

Furthermore, we analyze the numbers of CB-Patterns discovered in the synthetic dataset under different parameter settings, as shown in Figure 8. The number of the CB-Patterns is maximized at approximately *EPS.d* = 60 km and *EPS.t* = 16 days. When *EPS.d* increases further, the number of patterns decreases considerably; this number changes only slightly when *EPS.t* varies from 12 days to 16 days. This outcome indicates that a desirable *EPS.d* value is roughly 60 km and that a suitable setting for *EPS.t* is approximately 12 days. Therefore, our results are consistent with the spatiotemporal Gaussian distribution in the data generator.

5. Discussion

In this article, we investigated the application of the proposed CB-Pattern through the sequential behaviors exhibited by animals during migration. This pattern is an extension of the spatiotemporal sequential pattern: regular and repeated migration behaviors are derived in the context of their spatiotemporal attributes through the described automated process. In addition, similarities are detected in multiple individual movements and the process of translating these similarities into population-level patterns is determined. As illustrated through the wild bird and synthesized datasets, the CB-Pattern can tolerate a certain degree of divergence and perturbation among real-life similar migration behaviors. On the basis of such patterns, we can infer the behavioral states of specific species, and of different functional areas for the various stages of animal life.

As a particular species transitions between distinct annual life cycle states, generalized patterns of movement and behavior are recognizable, and therefore may be distinguishable based on these patterns. These patterns can also be extended to special tracking scenarios in which regional movements are a result of behavior states. The fundamental definition of representative spot permits two or more spots to coexist and overlap in time. These spots are combined to represent a generalized behavioral state owing to the adoption of the early selection principle (Palma *et al.* 2008) in the CGAR framework. Continuous movements between two areas for foraging and feeding are

extracted as a common spot in CGAR. If necessary, we can retain multiple overlapped spots by modifying the spot extraction process in CGAR and applying the distance to the first point in the spot as the distance constraint. We can also consider these moves to be multiple repeating spots by tuning gap threshold δ_G in the spot definition. Behavioral states for sufficient duration can also be implemented by incorporating an additional constraint into the spot definition.

The five typical CB-Pattern sequences for bar-headed geese and ruddy shelducks in the Tibetan Plateau region highlighted the ecological requirements for multiple stop-over and staging areas as sought by species during migration periods and for the linkage of breeding and wintering grounds during both migration stages. Likewise, the confluence of the two breeding populations from Qinghai Lake and Mongolia was clearly identified in this process, thus highlighting a region of potential population exchange and potential bottleneck to the species if environmental conditions resulted in a density-independent impact to survival (Takekawa et al. 2009, Prosser et al. 2011). Finally, the collective analysis conducted across marked individuals through the application of the CB-Pattern process on high-frequency and high-accuracy datasets may help highlight and identify the movement differences among individuals to the population level. Perhaps more importantly, by generating an inventory of sites identified by the CB-Pattern, the current analysis creates baseline information for future comparison to detect changes in annual site fidelity patterns or in habitat availability at certain sites as a result of environmental or anthropogenic influences. This information can guide specific conservation efforts in different areas, different life stages, and for different species.

6. Conclusion

In this article, we propose a notion of continuous behavior patterns to model the underlying common sequential behaviors during animal migrations. The continuous behavior pattern is defined as a sequence of common behaviors shared by multiple individuals, and is derived from individual similar movement fragments. Common behaviors are identified based on the spatiotemporal clustering of movement phases. The behavioral states of the overall pattern can be inferred depending on the spatiotemporal and sequential contexts. We also design a CGAR framework to derive all continuous behavior patterns from a raw trajectory dataset, and further propose an intuitively processing strategy to improve the framework performance. The results of experiments performed on three datasets comprising wild birds that migrate internationally show that our approaches are practical for discovering and portraying the spatiotemporal semantics of collective animal migration behaviors. Finally, we assess the running cost of the pattern discovery framework by conducting extensive experiments and a comprehensive analysis based on synthetic datasets. This work aids biologists in understanding collective animal activities, in determining the similarities among individuals or populations, and in identifying different behavioral stages and associated functional areas in an ecologically insightful manner. This article also presents a potential approach for the study of the relationships among animals according to trajectories collected during different periods. Finally, this study can provide useful information to guide specific conservation efforts in different areas, life stages, and for different species.

Note

1. http://www.werc.usgs.gov/ResearchTopicPage.aspx?id=17.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Alerstam, T. and Lindström, Å., 1990. Optimal bird migration: the relative importance of time, energy, and safety. *In*: E. Gwinner, ed. *Bird migration*. Berlin: Springer, 331–351.

Altizer, S., Bartel, R., and Han, B.A., 2011. Animal migration and infectious disease risk. *Science*, 331 (6015), 296–302. doi:10.1126/science.1194694

Benkert, M., *et al*., 2008. Reporting flock patterns. *Computational Geometry*, 41 (3), 111–125. doi:10.1016/j.comgeo.2007.10.003

Black, J.M., Prop, J., and Larsson, K., 2007. *Wild goose dilemmas: population consequences of individual decisions in barnacle geese*. Groningen: Branta Press.

Bridge, E.S., *et al*., 2011. Technology on the move: recent and forthcoming innovations for tracking migratory birds. *BioScience*, 61 (9), 689–698. doi:10.1525/bio.2011.61.9.7

Cao, H., Mamoulis, N., and Cheung, D.W., 2005. Mining frequent spatio-temporal sequential patterns. *In*: *Proceedings of the fifth IEEE international conference on data mining*. Houston, TX: IEEE.

Cui, P., *et al*., 2011. Movement patterns of bar-headed geese *Anser indicus* during breeding and post-breeding periods at Qinghai Lake, China. *Journal of Ornithology*, 152 (1), 83–92. doi:10.1007/s10336-010-0552-6

Drent, R., *et al*., 2003. Pay-offs and penalties of competing migratory schedules. *Oikos*, 103, 274–292. doi:10.1034/j.1600-0706.2003.12274.x

Ester, M., *et al*., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In*: *Proceedings of the 2nd international conference on knowledge discovery and data mining*. Menlo Park, CA: AAAI Press.

Faaborg, J., *et al*., 2010. Conserving migratory land birds in the new world: do we know enough? *Ecological Applications*, 20 (2), 398–418. doi:10.1890/09-0397.1

Fretwell, S.D. and Lucas, H.L., 1969. On territorial behavior and other factors influencing habitat distribution in birds. *Acta Biotheoretica*, 19, 16–36. doi:10.1007/BF01601953

Giannotti, F., *et al*., 2007. Trajectory pattern mining. *In*: *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM.

Jeung, H., *et al*., 2008. Discovery of convoys in trajectory databases. *Proceedings of the VLDB Endowment*, 1 (1), 1068–1080. doi:10.14778/1453856

Johnson, D.H., 1980. The comparison of usage and availability measurements for evaluating resource preference. *Ecology*, 61, 65–71. doi:10.2307/1937156

Li, Q., *et al*., 2008. Mining user similarity based on location history. *In*: *Proceedings of the 16th ACM SIGSPATIAL international conference on advances in geographic information systems*. New York: ACM.

Li, Z., *et al*., 2011. MoveMine: mining moving object data for discovery of animal movement patterns. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2 (4), 37.

Newton, I., 2008. *The migration ecology of birds*. London: Academic Press.

Owen, M., 1980. *Wild geese of the world: their life history and ecology*. London: B.T. Batsford.

Palma, A.T., *et al*., 2008. A clustering-based approach for discovering interesting places in trajectories. *In*: *Proceedings of the 2008 ACM symposium on applied computing*. Fortaleza: ACM.

Pedersen, M.W., *et al*., 2011. Estimating animal behavior and residency from movement data. *Oikos*, 120 (9), 1281–1290. doi:10.1111/more.2011.120.issue-9

Pei, J., *et al*., 2001. Prefixspan: mining sequential patterns efficiently by prefix-projected pattern growth. *In*: *Proceedings of the 17th international conference on data engineering (ICDE 2001)*. Heidelberg: IEEE Computer Society.

Prop, J., Black, J.M., and Shimmings, P., 2003. Travel schedules to the high arctic: barnacle geese trade-off the timing of migration with accumulation of fat deposits. *Oikos*, 103, 403–414. doi:10.1034/j.1600-0706.2003.12042.x

Prosser, D.J., *et al*., 2011. Wild bird migration across the Qinghai-Tibetan plateau: a transmission route for highly pathogenic H5N1. *Plos One*, 6 (3), e17622. doi:10.1371/journal.pone.0017622

Prosser, D.J., *et al*., 2009. Satellite-marked waterfowl reveal migratory connection between H5N1 outbreak areas in China and Mongolia. *Ibis*, 151 (3), 568–576. doi:10.1111/j.1474-919X.2009.00932.x

Pukelsheim, F., 1994. The three sigma rule. *The American Statistician*, 48 (2), 88–91.

Takekawa, J., *et al*., 2009. Geographic variation in bar-headed geese *Anser indicus*: connectivity of wintering areas and breeding grounds across a broad front. *Wildfowl*, 59 (59), 100–123.

Takekawa, J.Y., *et al*., 2010. Victims and vectors: highly pathogenic avian influenza H5N1 and the ecology of wild birds. *Avian Biology Research*, 3 (2), 51–73. doi:10.3184/175815510X12737339356701

Tang, L.-A., *et al*., 2013. A framework of traveling companion discovery on trajectory data streams. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5 (1), 3.

Trasarti, R., *et al*., 2011. Mining mobility user profiles for car pooling. *In*: *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*. San Diego, CA: ACM.

Wang, Y., *et al*., 2014. Mining continuous activity patterns from animal trajectory data. *In*: *Advanced data mining and applications*. Guilin: Springer, 239–252.

Zhang, X., *et al*., 2004. Fast mining of spatial collocations. *In*: *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*. Seattle, WA: ACM.