DATA MINING APPROACHES FOR HABITATS AND STOPOVERS DISCOVERY OF MIGRATORY BIRDS

XU Qiang[12], LUO Ze[1*], WEI Ying[12] and YAN Baoping[1]

[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China
[2]Graduate University of the Chinese Academy of Sciences, Beijing, 100190, China
Email: xuqiang@cnic.cn
*[*1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China*
Email: luoze@cnic.cn
[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China
[2]Graduate University of the Chinese Academy of Sciences, Beijing, 100190, China
Email: weiying@cnic.cn
[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China
Email: ybp@cnic.cn

ABSTRACT

This paper mainly focuses on using data mining technology to efficiently and accurately discover habitats and stopovers of migratory birds. The three methods we used are as follows: 1. Density-based clustering method, detecting stopovers of birds during their migration through density-based clustering of location points. 2. Location histories parser method, detecting areas that have been overstayed by migratory birds during a set time period by setting time and distance thresholds. 3. Time-parameterized line segment clustering method, clustering directed line segments to analysis shared segments of migratory pathways of different migratory birds, and discovers the habitats and stopovers of these birds. At last, we analyzed the migration data of bar-headed goose in the Qinghai Lake Area through the three methods above and verified the effectiveness of the three methods, and by comparison, identified the scope and context of use of these three methods respectively.

Keywords: Migratory birds; Flyway; Satellite tracking data; Detection algorithm; Bar-headed goose

1 INTRODUCTION

One of the most important tasks to protect migratory birds around the globe is to identify the ecological needs of birds in their breeding and wintering grounds as well as the stopovers during their migration (Berthold, &Terrill, 1991). The information of specific migration routes, net structures of these migration routes and important stopovers during migration is the key to research migratory birds' selection of habitats and stopovers, birds' migration strategy and the influence of global climate change on migratory birds' migration. On the other hand, the role of migratory birds in the spread of avian influenza virus has been a hot topic nowadays. Among the wild birds which have been infected by the H5N1 highly pathogenic avian influenza virus, many are migratory, so migratory bird might be avian influenza virus vectors. As the ecological environment and natural resources of the habitats and stopovers might set the stage for interspecific or intraspecific transmission of avian influenza virus among birds, studying wild birds' migration and detecting these birds' habitats or stopovers efficiently and precisely are of significant value for the research and prevention of the spread of avian influenza virus.

The traditional way of studying bird migration, like bird banding, is simple and easy to carry out, but its result depends on long-time observation and the number and quality of returned birds are under expectation, thus it's impossible to get a whole picture of the track of bird migration in short time (Zhang, &Yang, 1997). In other words, the traditional way is hard to meet the requirements of modern study. The development of satellite tracking technology and its application in biology in recent years provide new opportunities for bird migration study (Cagnacci, Boitani, Powell, & Boyce, 2010). Some of the raw data by using satellite tracking technology is shown in the following **Table 1**.

Table 1. Relational representation of raw GPS data.

ID	Animal	Latitude	Longitude	lc94	Date time
930796	BH07_67582	65.448	96.317	LZ	2008-01-30 04:02:00
930948	BH07_67582	65.448	96.317	LZ	2008-01-30 04:02:00

In this chart, **ID** is the recording number, **Animal** is the label of the migratory bird, **Latitude** and **Longitude** showing the specific location, and the **Date time** field signifying time stamp. Obviously，traditional data analysis methods such as drawing-dot or manual statistics method cannot process these high-resolution spatial-temporal data. This paper mainly focuses on using data mining technology to discover habitats and stopovers of migratory birds among the original satellite telemetry data efficiently and accurately, these methods are described as follows:

- **Density-based clustering method**. The habitats and stopovers of migratory bird are the areas where the bird continuously stays for some time, corresponding to the dense regions in space. We use the density–based clustering method to discover these dense regions. Although the location data of the migratory bird may be lost because of some different reasons, these dense regions can characterize the habitats or stopovers of the bird.
- **Location histories parser method**. Given a time and distance threshold, modeling the move status (stay or move) of migratory bird, and then scanning a certain bird's migration route point by point. This method can get the arriving and leaving time of the migratory bird at its every stopover.
- **Time-parameterized line segment clustering method**. We measure the space-time density of moving objects by the spatial distance, the direction of the movement and the time characteristics. We use the time-based plane-sweeping trajectory clustering algorithm to analysis shared segments of migratory pathways of different migratory birds, and discover the habitats and stopovers of these birds.

The following part of this paper is organized as the following: the second section introduces some relevant researches; the third section defines some specific terms; the forth section elaborates three ways to discover stopovers among GPS data; the fifth section presents the experiments and the result analysis; the last section provides the major conclusions of the paper.

2 RELATED WORK

As the improvement in GPS-based radio telemetry and growing international concern about the migratory birds, many international organizations began to trace the birds' migration through satellite positioning technology (Frisch, Vagg, & Hepworth, 2006). There is increasing interest on developing methods to perform data analysis for trajectory datasets (Schiller, &Voisard, 2004) (Stauffer, &Grimson, 2000). A typical data analysis task is to detect the stopovers of the moving objects. We used the same satellite telemetry datasets with (Tang et al.,2009), Tang et al. (2009) proposed a hierarchical spatial clustering method HDBSCAN to find the habitats or stopovers of migratory birds in different spatial scale levels, but HDBSCAN algorithm measures the proximity of birds mainly by Euclidean distance between two points and does not take time information into account. Hariharan, & Toyama(2004), Zheng, Zhang, Ma, Xie, &Ma(2011), Zheng, &Li(2008), Zheng, &Xie(2010) modeled the location histories of human and proposed a method to find the stopovers of human, but their attention focused on personalized recommendation based on location, so they did not study the stopovers in depth. Gaffney, & Smyth(1999), Gaffney, Robertson, Smyth, Camargo, & Ghil(2006) observed that existing trajectory clustering algorithms group similar trajectories as a whole, thus revealing common trajectories. But clustering trajectories as a whole could not detect similar portions of trajectories or could miss common sub-trajectories. The framework and algorithm proposed by Lee, Han, & Whang (2007) did not consider temporal information. Satellite telemetry datasets or GPS-based locations datasets are essentially time series of spatial data. To measure the space-time density of moving objects, this paper defines different distance functions from (Lee et al., 2007) to measure the similarity of different line segments, so that we can find the shared segments of migratory pathways both in time and space. In this paper, we use three data mining methods to discover habitats and stopovers of migratory birds, and analyze in detail the characteristics and the contexts of use of the three algorithms respectively.

3 PRELIMINARY

In this section, we clarify some terms used in this paper such as **point**, **line segment**, **trajectory** etc.

Point: a point P is indicated by a tuple$< Lat, Lng >$, which refers to that one bird once presented in a location at where the latitude is *Lat* and the longitude is *Lng*.

Point set: a point set *PS* consists of a series of points which are generated by one or more birds.

Trajectory: a trajectory *TR* is defined as an ordered set of $< position, timestamp >$ pairs ordered by time serials. $TR = \{< P_1, t_1 >, < P_2, t_2 >, < P_3, t_3 >, \ldots, < P_n, t_n >\}$, $\forall(i < j)$ $t_i < t_j$ where t_i is point P_i 's timestamp.

Line segment: Given a trajectory *TR*, a line segment of *TR* is defined as $LS_i =<< P_i, t_i >, < P_{i+1}, t_{i+1} >>$, where $< P_i, t_i >, < P_{i+1}, t_{i+1} > \in TR$ represents object moves from position P_i to position P_{i+1} during $[t_i, t_{i+1}]$. The displacement of moving object is denoted by $\overline{LS_i}$, and the duration of LS_i is denoted by $LS_i.TD$.

Line segment set: The line segment set of a trajectory *TR* is defined as a collection of two sequential pairs in *TR*, $LSS = \{< P_i, t_i >, < P_{i+1}, t_{i+1} > | 1 \leq i \leq n-1\}$

Stop region: stop region is the area where the migratory birds stay for some time during their migration. Migratory birds' habitats and stopovers are all stop regions. We use a stop region center's coordinate to indicate the stop region in the following sections.

4 THREE METHODS TO DISCOVER THE STOP REGIONS

Migratory routes of migratory birds are long and complex paths (**Figure 1**), and the migratory birds' raw GPS data can't be used conveniently due to its large scale and high complexity. In this section, we will provide three methods to solve the problem, and explain their principle in detail.

Figure 1. Migratory pathway of one bar-headed goose captured in the Qinghai Lake Area.

4.1 Density-based clustering method

As depicted in **Figure 1**, the dense regions in the picture may be the stop regions from the visual point of view. We can assume that dense regions in spatial-temporal data are equivalent to the stop regions. The GPS position sampling frequency of satellite telemetry device was about once every 2 hours during the day. If a bird stays in a small area more than a certain period of time, the sampling point in this area may be denser than other place. So it is possible to detect the migratory birds' stop regions by finding the dense areas in GPS location history data.

In order to find the dense clusters in spatial data, Ester, Kriegel, Sander, & Xu(1996) proposed the DBSCAN algorithm. The density-based algorithm based on the following notions: ***ε*-neighborhood** is the neighborhood within a radius ε of a given object; an object is a **Core object** if the ***ε*-neighborhood** of this object contains at least a minimum number (*MinPts*) of objects; an object p is **directly density-reachable** from object q if p is within the ***ε*-neighborhood** of q, and q is a core object; an object p is **density-reachable** from object q with respect to ε and *MinPts* in a set of objects, *D*, if there is a chain of objects $p_1, \ldots, p_n$, where $p_1 = q$ and $p_n = p$ such that p_{i+1} is **directly density-reachable** from p_i with respect to ε and *MinPts*, for $1 \leq i \leq n, p_i \in D$; an object p is **density-connected** to object q with respect to ε and *MinPts* in a set of objects, *D*, if there is an object $o \in D$ both p and q are **density-reachable** from o with respect to ε and *MinPts* (Han, &Kamber, 2000). All points within the cluster are mutually density-connected. If a point is density-connected to any point of the cluster, it is part of the cluster as well.

The stop region detection algorithm based on DBSCAN (Ester et al., 1996) is described as follows:

Input: Point set: *PS*; Radius: ε; the minimum number of points to decide the core objects: *MinPts*
Out Put: A set of all stop regions *SS*
DBS_SR_DETECTION (*PS*, ε, *MinPts*):
    C = 0;
    For each unvisited point *P* in dataset *PS*
        Mark *P* as visited;

```
        N = P's ε-neighborhood set;
        If P is not a core object
            Mark P as NOISE;
        Else
            C++;
            Add P to cluster C;
            For each point O in N        //find all the objects that density-connected with P
                If O is not visited
                    Mark O as visited;
                    N' = O's ε-neighborhood set;
                    If O is a core object
                        N = N joined with N';
                If O is not yet member of any cluster
                    Add O to cluster C;
    Return the center coordinates of each cluster;
```

The time complexity of *DBS_SR_DETECTION* is $O(n^2)$, where n is the number of points in *PS*. If the appropriate spatial index is used, the time complexity of this algorithm will reduce to $O(n\log n)$. If ε and *MinPts* are appropriately set, this algorithm can detect arbitrarily shaped clusters, but as for how to choose these two parameters there is no good way. When we use this algorithm, *PS* can be either one bird's history location set or multi-birds' history location set. Here, $PS = \{P_1, P_2, .., P_n\}$, where $P_i =< Lat_i, Lng_i >$, the points in *PS* only contain spatial dimension, and we use great-circle distance as geographical distance formula between two points. Furthermore, the NOISE in *DBS_SR_DETECTION* may be significant for the ornithologist, because the object may be flying fast at this location.

4.2 Location histories parser method

As stated before, the *DBS_SR_DETECTION* only takes the spatial dimension into account, dismissing the time dimension. In fact, birds' migration routes are complex and not regular (**Figure 1**), and the bad climate or other factors in wild environment may cause satellite signal lost. As depicted in **Figure 2**, if we use *DBS_SR_DETECTION* algorithm to detect this bird's stop region, we may find out the region surrounded by dotted red line, obviously showing that region is meaningless.

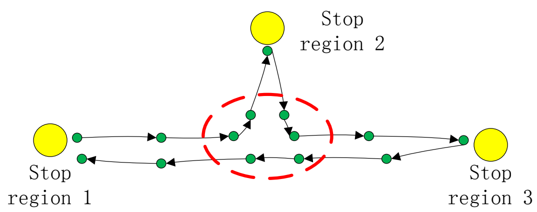

Figure 2. A typical migratory route

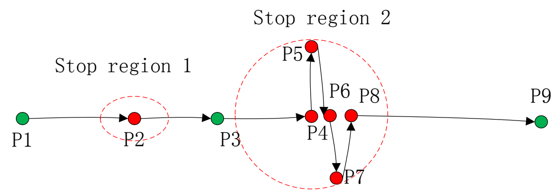

Figure 3. Two kinds of stops of migratory birds

In order to solve the problem above, we need take the time dimension into account. Hariharan et al.(2004),Zheng et al.(2011),Zheng et al.(2008),Zheng et al.(2010) proposed a time and distance threshold based method to discover human's stay point from the historical location data. This method may be useful for detecting the migratory birds' stop regions. The stops of migratory birds may be divided into two kinds:

- As the stop region 1 depicted on **Figure 3**, during the migration, birds may keep stationary for some time because of the bad weather or they need a rest.
- As the stop region 2 depicted on **Figure 3**, the birds may stay in a little area for some time, because they need to find food or for some other reasons.

Both of the stops can be defined as this:

Given a trajectory $TR = \{< P_1, t_1 >, < P_2, t_2 >, < P_3, t_3 >, \ldots, < P_n, t_n >\}$, if there is a subset of TR $sTR =$

$\{< P_i, t_i >, < P_{i+1}, t_{i+1} >, \ldots, < P_j, t_j >\}$ where $1 \leq i, j \leq n$ and for $\forall i \leq k \leq j$, $\mathrm{Dist}(P_i, P_k) \leq Dr$, $\mathrm{Dist}(P_i, P_{j+1}) > Dr$, $\mathrm{Int}(t_i, t_j) \geq Tr$, the $\mathrm{Dist}(P_i, P_k)$ denotes the geospatial distance between two points P_i and P_k, the $\mathrm{Int}(t_i, t_j) = |t_i - t_j|$ is the time interval between two points, then the area where the points at *sTR* locate is a stop region *S* (Zheng, &Xie, 2010). We can also use a quaternion to indicate a stop region $S =< Lat, Lng, ts, te >$. The *Lat* stands for the average latitude of the collection *sTR*; the *Lng* stands for the average longitude of the collection *sTR*; the *ts* means the bird's arriving time on stop region *S*; the *te* means bird's leaving time. We can compute them as: $S.Lat = \frac{\sum_{k=i}^{j} P_k.Lat}{|sTR|}$, $S.Lng = \frac{\sum_{k=i}^{j} P_k.Lng}{|sTR|}$, $S.ts = t_i$, $S.te = t_j$.

The algorithm that detects all stop regions from a trajectory is described as follows:

Input: A trajectory: *TR*; Distance threshold: *Dr*; Time threshold: *Tr*
Output: A set of all stop regions *SS*
LHP_SR_DETECTION (*TR*, *Dr*, *Tr*):
    i=0, $n = |TR|$;   //the number of GPS points in a GPS logs
    While *i* < *n* do:
        j=*i*+1;
        While *j* < *n* do:
            $Dist = \mathrm{Dist}(P_i, P_j)$
            If $Dist > Dr$ then
                $\Delta T = \mathrm{Int}(t_i, t_j)$;
                If $\Delta T > Tr$ then
                    $S.Lat = \sum_{k=i}^{j} P_k.Lat/(j - i + 1)$;
                    $S.Lng = \sum_{k=i}^{j} P_k.Lng/(j - i + 1)$;
                    $S.ts = t_i$; $S.te = t_j$;
                    SS.insert(*S*);
                i=*j*+1; break;
            j=*j*+1;
    Return *SS*

This algorithm's time complexity in the worst case is $O(n^2)$. The data *LHP_SR_DETECTION* can process is one bird's trajectory. Before use this algorithm we should sort the bird's location history data by timestamp. This algorithm can't deal with multi-birds' trajectory. A simple method to solve this problem is to combine *DBS_SR_DETECTION* with *LHP_SR_DETECTION*, which can detect all stop regions of one bird respectively, and then cluster all the stop regions of all birds.

4.3 Time-parameterized line segment clustering method

Birds in the same region usually share their habitats or stopovers. As indicated in **Figure 4**, different birds fly from one same place to another, and as a result many similar line segments will be generated between these two places. The sets of starting points and finishing points of each line segments in this cluster may be the stopovers or habitats of migratory birds.

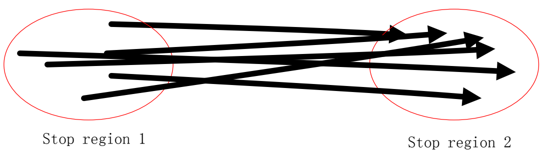

Figure 4. A line segment cluster

In order to cluster line segments, the first problem need to be solved is to measure the distance between two objects. The distance function we proposed to measure the distance between two line segments includes both spatial and temporal aspects. We define the distance function between line segments $LS_i =<< P_i, t_i >, <$

$P_{i+1}, t_{i+1} >>$ and $LS_j = << P_j, t_j >, < P_{j+1}, t_{j+1} >>$ as follow:

$$\text{L\_dist}(LS_i, LS_j) = \begin{cases} \frac{\text{dist}(P_i, P_j) + \text{dist}(P_{i+1}, P_{j+1})}{2}, \text{ if } LS_i.TD \cap TW \neq \emptyset \wedge LS_j.TD \cap TW \neq \emptyset \wedge \angle(LS_i, LS_j) \leq \Theta \\ \varepsilon + 1, \text{else} \end{cases} \quad (1)$$

Here ε means the spatial threshold; θ means the angle threshold; $\text{dist}(P_i, P_j)$ means the distance between two points P_i and P_j, the distance is measured by the great circle distance; $\angle(LS_i, LS_j)$ means the included angle between line segments LS_i and LS_j, which is measured by the spherical angle between two great circles containing line segments; *TW* means the time window $TW = [t_1, t_2]$.
After defining the distance function between two line segments, we use the DBSCAN(Ester et al., 1996) algorithm to find all the dense clusters. As the object we concerned is line segment, we give some extra description. The set of all the line segments is denoted as *LSC*; the ε-neighborhood set of line segment $LS_i(LS_i \in LSC \wedge LS_i.TD \cap TW \neq \emptyset)$ in time window *TW* is defined as:

$$N_{(\varepsilon,TW)}(LS_i) = \{LS_k | LS_k \in LSC \wedge LS_k.TD \cap TW \neq \emptyset \wedge \text{L\_dist}(LS_i) \leq \varepsilon\} \quad (2)$$

The algorithm can be described as follows:

Input: The set of all line segments: *LSC*; the time window: *TW*; Distance threshold: ε; Minimum number of line segments: *MinLSSum*; the angle threshold: θ
Output: A set of stay region *SS*

```
TPLS_SR_DETECTION (LSC, TW, ε, MinLSSum, θ):
    LSC_new= {}     //get rid of the NOISE in advance
    For each line segment LS in LSC
        If LS.TD ∩ TW ≠ ∅
            LSC_new .add(LS);
    C = 0;
    For each unvisited line segment LS in dataset LSC_new
        Mark LS as visited;
        N = N_(ε,TW)(LS);
        If Size of (N) < MinLSSum
            Mark LS as NOISE;
        Else
            C++;
            Add LS to cluster C;
            For each line segment LS' in N
                If LS' is not visited
                    Mark LS' as visited;
                    N' = N_(ε,TW)(LS');
                    If Size of (N') >= MinLSSum;        //if LS' is a core object
                        N = N joined with N';
                If LS' is not yet member of any cluster
                    Add LS' to cluster C;
    Return SS;        //get the set of all stay regions
```

The time complexity of the algorithm above is $O(n^2)$, where *n* is the number of the line segments in *LSC_new*. If spatial index is used, the time complexity will reduce to $O(n \log n)$. The algorithm *TPLS_SR_DETECTION* can only detect stop regions where the birds leave or arrive at during *TW*. In order to find all stop regions, we use the time window size *TWS* and time step *ts* to replace time window *TW* where *ts*<<*TWS*. Given a set of line segments *LSC*, *startTime* means the time of first location in *LSC*, *endTime* means the time of last location in *LSC*. A set of time window:

$$TW_{\text{set}} = \{[startTime, startTime + TWS],\ [startTime + ts, startTime + ts + TWS], [startTime + 2 * ts, startTime + 2 * ts + TWS], \ldots, [startTime + n * ts, endTime]\} \quad (3)$$

We use the time window parameter in TW_{set} respectively to call the function *TPLS_SR_DETECTION*, and then merge all the results. If the time window size and time step are appropriately set, we can detect all the stop regions. More details are described as follows:

Input: The set of all line segments: *LSC*; the time window size: *TWS*; time step: *ts*; Distance threshold: ε; Minimum number of line segments: *MinLSSum*; the angle threshold: θ
Output: A set of all the stay regions *SS*
TPLS_ALL_SR_DETECTION (*LSC*, *TWS*, *ts*, ε, *MinLSSum*, θ):
Sort *LSC* by time;
startTime=*LSC*.getStartTime (); //get the first location's time stamp
endTime=*LSC*.getEndTime (); //get the last location's time stamp
Get the set of the time window TW_{set};
SS={};
For each time window *TW* in TW_{set}
SS_{TW} = TPLS_SR_DETECTION($LSC, TW, \varepsilon, MinLSSum, \theta$);
$SS = SS \cup SS_{\text{TW}}$;
Return *SS*;

---

5 EXPERIMENTAL EVALUATION AND RESULT ANALYSIS

To verify the efficiency of these three methods, we choose the satellite telemetry data obtained from 29 bar-headed geese captured in the Qinghai Lake Area to run a series of tests. Raw data included 471,774 records of position and time information between 25 March 2007 and 5 June 2009. We selected 40,756 records with higher precision estimates to improve the reliability of analysis.

For *DBS_SR_DETECTION*, *PS* is the location history obtained from a bar-headed goose numbered BH07_74901, which has 3502 records of time and location information between 31 March 2007 and 23 November 2008. Under the condition of $\varepsilon = 20\text{Km}, MinPts = 10$, we find 11 stop regions during this bird's migration (**Figure 5**).The distribution of the stop regions we detected are indicated as **Table 2**.

For *LHP_SR_DETECTION*, *TR* is the trajectory obtained from the same bar-headed goose as above. Under the condition of $Dr = 20\text{Km}, Tr = 48\text{h}$, we find 31 stop regions (**Figure 6**). These 31 stop regions are distributed as **Table 2**.

For *TPLS_ALL_SR_DETECTION*, the GPS position sampling frequency of the satellite telemetry device was about once every 2 hours during the day. We reduced the dimension of data from hours to days by choosing 2 positions that spanned two sampling times closest to a day. These two locations were regarded as starting and ending points of a line segment. The duration between two sampling times was the duration *TD* of the line segment. At last we choose 5,959 line segments to make up *LSC*. The time interval is from 25 March 2007 to 4 June 2009. Under the condition of *TWS*=60days, $\varepsilon = 80\text{Km}$, $MinLSSum = 2$, $\theta = 10$ degrees. Detailed results are as **Figure 7**. The stop regions we detected are: Qinghai Lake Area; The river valleys near Lhasa; Eling Lake and Zaling Lake; Niriacuogai Lake, Zamucuo Lake, and Gaeencuonama Lake; and Cuona Lake, Cuoe Lake, Nairipingcuo Lake.

Table 2. The distribution of stop regions generated by *DBS_SR_DETECTION* and *LHP_SR_DETECTION*

Area	**Stop region (*DBS_SR_DETECTION*)**	**Stopregion (*LHP_SR_DETECTION*)**
Qinghai Lake Area	Stop region 9,10	Stop region 1,2,3,4,5,6,7
DonggeiCuona Lake Area	Stop region 11	Stop region 8,9,10
Eling Lake and Zaling Lake Area	Stop region 7	Stop region 11,12
Galalacuo Lake Area	Stop region 8	Stop region 13
Saiyongcuo Lake Area	Stop region 5	Stop region 21,22,23
Zhamucuo,Niri'a cuogai,Ga'e Encuo Nama Area	Stop point 6	Stop point 24,25,26,27,28,29
Cuo'e Lake and Neri puncuo Area	Stop region 2	Stop region 14,15,17,18,19,30
River valleys near Lhasa	Stop region 1	Stop region 16,31

Figure 5. The white mark means NOISE, the marks with the same color belonging to one stop region; the number near the mark is the stop region number.

Figure 6. A yellow mark is a stop region and the number near the mark is the stop region number.

Figure 7. Clustering results of long distance segments from 12 June 2007 to 2 April 2009.

From the results above, we can figure out that the stop regions obtained by executing *DBS_SR_DETECTION* and *LHP_SR_DETECTION* to analyze the same bird's migratory route are similar to each other. Nearly all the stop regions are next to lakes or wet lands (**Figure 5**, stop region 5, stop region 7, stop region 11). While the data handled by *TPLS_ALL_SR_DETECTION* are from all bar-headed Geese, it's not suitable to be compared with the other two algorithms. But we still can find that stop regions obtained by these three algorithms have obvious overlapping areas. Moreover, the result of *TPLS_ALL_SR_DETECTION* is almost the same as the stop regions of bar-headed Geese's migratory routes mentioned in (Tang et al., 2009).

The distance thresholds of *DBS_SR_DETECTION* and *LHP_SR_DETECTION* are both 20Km, while stop regions obtained from *LHP_SR_DETECTION* are much more than those from *DBS_SR_DETECTION*. Based on these two algorithms' principles, *DBS_SR_DETECTION* only considers the information of spatial dimension, so we can only find out its dense clusters and treat them as stop regions. From microcosmic view, this algorithm is unable to analyze data within dense clusters. For instance, **Figure 8(a)** and **Figure 8(b)** indicate the same area.

DBS_SR_DETECTION treats this area as one stop region, while *LHP_SR_DETECTION* obtains several stop regions for it considers both spatial and time dimension. Although these stop regions' spatial positions are next to each other, treating them as different regions still means a lot. What's more, stop region 3(**Figure 8(c)**) discovered by *DBS_SR_DETECTION* is treated as noise (**Figure 8(d)**) when executing *LHP_SR_DETECTION*. We figure out that birds have 13 position points in the area but never stop there beyond one day and it is probably an exception for this area isn't a perfect stop region. However, *DBS_SR_DETECTION* still takes the area as a stop region while *LHP_SR_DETECTION* can avoid this incorrect situation. We also notice that stop regions detected by *DBS_SR_DETECTION* are without the information of time, while stop regions obtained by *LHP_SR_DETECTION* are ordered by time sequence. The three algorithms' further comparison is as **Table 3**.

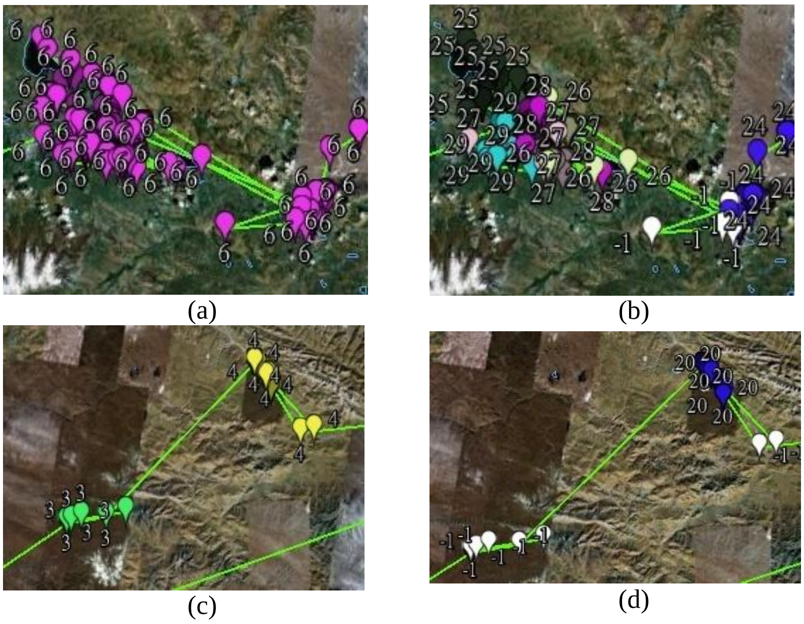

Figure 8. Four special scenarios, (a) and (c) are generated by *DBS_SR_DETECTION*, (b) and (d) are generated by *LHP_SR_DETECTION*

Table 3. Comparison of three methods in detail

	DBS_SR_DETECTION	***LHP_SR_DETECTION***	***TPLS_ALL_SR_DETECTION***
Dimension	Spatial	Spatial and time	Spatial and time
Object	Point	Trajectory	Line segment
Raw data	GPS location history	GPS location history	GPS location history
Range	One bird or more	One bird	Multiple birds
Time complexity	$O(n^2)$or $O(n\log n)$	$O(n^2)$	$O(n^2)$or $O(n\log n)$

From the experiments above, we find that all these three methods can detect habitats and stopovers on the bar-headed geese's migratory routes. However, their principles lead to their differences in application. *DBS_SR_DETECTION* does well in the situation that only cares about stop regions' position. For example, sometimes ornithologists need to know the common stopovers for the whole flock of bar-headed geese during their migratory routes. *DBS_SR_DETECTION* may be very suitable for this situation above. The object handled by *LHP_SR_DETECTION* is the trajectory, so this algorithm can only deal with one bird's trace once. If we want to analyze more birds' information, we need to perform it multiple times before further processing. This algorithm takes the time factor into account. We can detect stop regions with start and end timestamps, which indicate some bird's arriving and leaving time in some area. This may be useful for studying the relationship between the flyways of migratory birds and climate. The object handled by *TPLS_ALL_SR_DETECTION* is line segment. This algorithm is meaningful only when many birds' trajectories are analyzed. The intermediate products during the process are line segment clusters. According to those clusters, we can easily figure out the fly distance among the stop regions. As indicated in **Figure 7**, observing the lengths of line segment clusters, we find that stop regions around Eling Lake and Zaling Lake are most bar-headed geese's start areas before their long journey. Departing from there, some of the birds make a pit-stop in Niriacuogai Lake, Zamucuo Lake and

Gaeencuoname Lake while others fly at one go until Cuona Lake, Cuoe Lake and Nairipingcuo Lake. This information may be useful for ornithologists to analyze birds' migration patterns.

6 CONCLUSION

Over all, we provide three methods based on data mining for detecting habitats and stopovers on the migratory routes from birds' GPS data. After applying the algorithms on the GPS data of bar-headed Geese captured in the Qinghai Lake region of China, we verify the algorithms' correctness. Having analyzed their principles and distinctions in detail, we give some suggestions about the applying situations of these three algorithms. It'll be helpful for ornithologists to find apposite algorithm for their work. In the future, we'll further study the climate, ecology and other factors in the stop regions on birds' migratory routes.

7 ACKNOWLEDGEMENTS

Funding was provided by the Natural Science Foundation of China under Grant No. 90912006; the Natural Science Foundation of China under Grant No. 61003138; The National R&D Infrastructure and Facility Development Program of China under Grant No. BSDN2009-18; Special Project of Informatization of Chinese Academy of Sciences in "the Eleventh 5-Year Plan", e-Science Application of Research on Resources, Disease Monitoring and Risk Assessment of Important Wild Birds in the Qinghai Lake Region under Grant No. INFO-115-D02; Special Project of Informatization of Chinese Academy of Sciences in "the Eleventh 5-Year Plan", Basic Databases of Joint Research Center of Chinese Academy of Sciences and the Qinghai Lake National Nature Reserve under Grant No. INFO-115-C01-SDB2-02; Fund of President of Chinese Academy of Sciences; Found of Director of Computer Network Information Center of Chinese Academy of Sciences; United States Geological Survey (Patuxent Wildlife Research Center, Western Ecological Research Center, Alaska Science Center, and Avian Influenza Program); the United Nations FAO, Animal Production and Health Division, EMPRES Wildlife Unit; National Science Foundation Small Grants for Exploratory Research (No. 0713027); and the Chinese Academy of Sciences (No. 2007FY210700, KSCX2-YW-N-063 and 2005CB523007). The use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the U.S. Government.

8 REFERENCES

Berthold, P., & Terrill, S.B., (1991) Recent advances in studies of bird migration. *Annual Review of Ecology and Systematics* 22, 357-378.

Cagnacci, F., Boitani, L., Powell, R.A., & Boyce, M.S., (2010) Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. *Phil. Trans. R. Soc*. B 365(1550), 2157-2162.

Ester, M., Kriegel, H.-P., Sander J., & Xu, X., (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd international conference on knowledge discovery and data mining*, Portland, OR, USA.

Fuller, M.R., Seegar, W.S., & Howey, P.W., (1995) The use of satellite systems for the study of bird migration. *Israel Journal of Zoology* 41(3), 243-252.

Frisch, H., Vagg, R., & Hepworth, H.(Eds.), (2006) Migratory Species and Climate Change: Impacts of a Changing Environment on Wild Animals. CMS Convention on Migratory Species of Wild Animals/UNEP, Bonn, Germany.

Gaffney, S., & Smyth, P., (1999) Trajectory Clustering with Mixtures of Regression Models. *Proc. 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, San Diego, California, USA.

Gaffney, S.J., Robertson, A.W., Smyth, P., Camargo, S.J., & Ghil, M., (2006) Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models. *Climate Dynamics* 29(4), 423-440.

Hariharan, R., & Toyama, K., (2004) Project Lachesis: Parsing and Modeling Location Histories. *Geographic Information Science Lecture Notes in Computer Science* 3234, 106-124.

Han, J., & Kamber, M., (2000) Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers
Lee, J.-G., Han, J. & Whang, K.-Y., (2007) Trajectory clustering: a partition-and-group framework. *Proceedings*

of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W., (2008) Mining User Similarity Based on Location History. ACM GIS '08, New York, NY, USA.

Schiller, J., & Voisard, A.(Eds.), (2004) Location-Based Services, Morgan Kaufmann

Stauffer, C., & Grimson, W. E. L., (2000) Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(8), 747–757.

Tang, M., Zhou, Y., Cui, P., Wang, W., Li, J., Zhang, H., Hou, Y., & Yan, B., (2009) Discovery of Migration Habitats and Routes of Wild Bird Species by Clustering and Association Analysis. *Computer Science* 5678, 288-301.

Zhang, Y.F., & Yang, R.L., (1997) Bird Migration Research of China, Beijing published in: China Forestry Publishing House

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

Zheng, Y., & Xie, X., (2010) Learning Location Correlation from GPS Trajectorie. 2010.Mobile Data Management (MDM), 2010 Eleventh International Conference, Kansas City, Missouri, USA.