Discovery of Migration Habitats and Routes of Wild Bird Species by Clustering and Association Analysis

MingJie Tang[1, 3], YuanChun Zhou[1], Peng Cui[2,3], Weihang Wang[1,3], Jinyan Li[4],
Haiting Zhang[1,3], YuanSheng Hou[5], and BaoPing Yan[1]

[1] Computer Network Information Center, Chinese Academy of Sciences
[2] Institute of Zoology, Chinese Academy of Sciences
[3] Graduate University of Chinese Academy of Sciences
[4] School of Computer Engineering, Nanyang Technological University
[5] Bureau of Qinghai Lake National Nature Reserve
100190 Beijing
{tangrock,supercat0325}@gmail.com, yczhou@sdb.cnic.cn,
cuipeng@ioz.ac.cn, jyli@ntu.edu.sg, zht@sdb.cnic.cn,
houyuanseng@163.com, ybp@cnic.cn

Abstract. Knowledge about the wetland use of migratory bird species during the annual life circle is very interesting to biologists, as it is critically important for conservation site construction and avian influenza control. The raw data of the habitat areas and the migration routes can be determined by high-tech GPS satellite telemetry, that usually are large scale with high complexity. In this paper, we convert these biological problems into computational studies, and introduce efficient algorithms for the data analysis. Our key idea is the concept of hierarchical clustering for migration habitat localization, and the notion of association rules for the discovery of migration routes. One of our clustering results is the Spatial-Tree, an illusive map which depicts the home range of bar-headed geese. A related result to this observation is an association pattern that reveals a high possibility of bar-headed geese's potential migration routes. Both of them are of biological novelty and meaning.

Keywords: Clustering, Sequence mining, Bird Migration, Habitat, Route, Scientific data, Qinghai Lake.

1 Introduction

The Asian outbreak of highly pathogenic avian influenza H5N1disease in poultry in 2003 and 2004 was unprecedented in its geographical extent, and its transmission to human beings showed an ominous sign of life-threatening infection [1]. Research findings indicate that the domestic ducks in southern China played a central role in the reproduction and maintenance of this virus, and wild birds may have contributed to the wide spread of the virus. This assumption had led to another question: how to define and identify the habitat, migration distance and time. Indeed, understanding of the species' habitat is critical for us to find the roots of the answers, like answers to how the wild life and domestic poultry intersect together, what is the possibility of H5N1 spilling over from the poultry sector into some wild bird species [2].

R. Huang et al. (Eds.): ADMA 2009, LNAI 5678, pp. 288–301, 2009.

The spatial data analysis on the specie's transmission coordinates together with their layered maps can be conducted by GIS (Geographic Information System) including ESRI'S ARC\INFO 7.1.2 and ArcView 3.1 (Research Institute, Inc., Redlands, California, U.S.A.) [5]. However, there has been lack of a persuasive way to identify the stop area of the species and the wintering areas. The situation becomes further complicated when the scientist come to lineate the migration routes from the accumulated data points. Therefore, a bird migration data analysis system is desired, by which data can be systematically analyzed, and knowledge patterns are subsequently available for deep biological studies. In this work, we address the following three problems which are arisen from the bird migration data analysis.

Discovery of Bird Habitat. The habitat range of an animal is defined as the area explored by this individual during its normal activities (i.e., food gathering, mating and caring for young, Burt 1943, Powell 2000). Understanding the factors that determine the spatial coverage and distribution of animals is fundamental not only to theoretical science, but also to real-life applications such as conservation and wildlife management decision makings [5].

Analysis on the Site Connectedness between Habitats. Site connectedness is a measure relating to the accessibility, for the migrating storks, of the site to its neighboring stay sites [6]. The sites with lower connectedness are considered to those at higher risk of being isolated from the migration route network.

Identification of Migration Routes. To help conserve species that migrate long distances, it is essential to have a comprehensive conservation plan that includes identification of migration routes. This information is of an added importance for many rare wild bird species [4].

Our computational approaches to these problems are integrated into a data mining system. It consists of four major components: data preprocessing, clustering, habitat range estimation, and association rules analysis. The function of the clustering component is to cluster the data points and meanwhile identify the candidates of the habitats. Intuitively, a potential habitat is a region where wild bird species prefer to stay a long time, and it mathematically corresponds to a dense region of points over the entire area. For this purpose, we propose a new hierarchical clustering algorithm which can find the habitats with different levels of densities. The component of habitat range estimation is aimed to determine the precise home range and time duration of the birds on top of the clustering results. As bird's migration between the habitats can be considered as a sequence pattern, we apply an existing sequence mining technique to discover interesting associations between the habitats. This is the goal of the association rules analysis. Besides, a visualization technique is developed for an easy view of the distribution of the bird habitats and migration routes which is helpful to gain more insights into findings. With this visualization tool, all of our results can be embedded into the Google Map (One web GIS from Google).

We have conducted a pilot experiment on a real-world database to evaluate our system. Our computational results on the bird habitat, site connectedness and migration route are interesting and have been confirmed to have biological novelty. These results would be useful in future for the scientists to estimate the risk of virus infection of wild birds from poultry or the other way around.

The main contributions of our work are summarized as follows: (i) A new hierarchical clustering algorithm is proposed, and it is used to discover bird habitats, (ii) Association analysis is introduced to reveal the site connectedness between habitat areas, and (iii) Bird migration routes are rigorously studied by sequence mining algorithms.

Paper organization. Section 2 presents a short background introduction to satellite tracking technologies for monitoring migration routes of wild bird species, and gives a brief overview to clustering and association mining algorithms. In Section 3, the telemetry bird migration data is described. Section 4 presents the overall diagram of our data mining system and describes the computational techniques in each component. Our computational results and their evaluation are presented in Section 5. In Section 6, we summarize our major contribution, and point out our future work.

2 Background and Related Work

2.1 Satellite Tracking of Wild Bird Species

Recent advances in the technology of satellite tracking have allowed researchers to continuously track the movements of individual birds over a broad spatial scale without conducting extensive field observations after the birds have been equipped with satellite transmitters. The applications of satellite tracking to bird migration studies have enabled considerable progress to be made with regard to elucidating the migration routes and stay sites of various migratory bird species, with important implications, for example, for conservation [6]. Traditionally, most of biologist have to count those location plots in a certain area and then utilize kernel model to calculate the home range of bird species [3,5,6]. Until recently, Hiroto Shimazaki1 *et al.* [6] proposed a method to examine the location data points based on the idea of clustering. At the first step, their method groups the location points with similar characteristics in approaching speed and departure speed by using the ISODATA algorithm [9]. And then the extent of stay sites is determined by specifying the area attainable by a bird moving speed. At last, they evaluate the site connectedness between stopover sites. However they do not make full use of the bird tracking data—features such as latitude and longitude have not been used to get the habitat range. The identification of the migration routes has not been touched either. As shown in the previous studies that satellite tracking is a powerful to monitor birds' migration behavior, and the data is valuable to make significant contribution to biological research, yet, to the best of our knowledge, it has been long lack of a data mining system capable of conducting systematic migration data analysis.

2.2 Overview to Clustering Algorithms

Clustering is an extensively studied topic in the machine learning and data mining field. A clustering algorithm refers to a method that subgroups a set of data points according to a distance or density metrics. Clustering analysis can be used as a standalong tool to get insight into the distribution of the data points in a data set, or can be used as a data preprocessing step for other types of data analysis. Various techniques have been explored for clustering spatial data sets. For instance, an improved k-medoid method, called CLARANS [12] was proposed recently. SNN [13] was also developed to cluster the earth science data. DBSCAN [10] and IncrDBSCAN [11] have been proposed to process the spatial data sets as well. Meanwhile, several hierarchical clustering approaches have been long investigated, including the agglomerative approach (eg. AGNES) and the divisive approach (eg. DIANA). Detailed description of AGNES and DIANA can be found at [14]. In this paper, our new idea is to combine DBSCAN with a hierarchical clustering approach to find the habitats with different levels of densities.

2.3 Association Analysis

Association rules mining and sequence mining are pioneer research topics in data mining, and they are still attracting lots of attentions. The classic association rule mining algorithms include Apriori[15] and FP-tree[16]. GSP [17] was the first approach to the discovery of frequent sequence patterns. Zaki then propose the SPADE algorithm [18] to find frequent sequence with a faster speed. The PSP (Prefix Tree For Sequential Patterns) approach [21] is much similar to the GSP algorithm, but it stores the database on a more concise prefix tree with the leaf nodes carrying the supports of the sequences. In this paper, we make use of these algorithms for bird migration routes analysis.

3 Bird Migration Data

Our studies are conducted at the Qinghai Lake National Nature Reserve, Qinghai province, China. Qinghai Lake, the largest salt lake in China with an area of 525 Km2, is located in the middle of Qinghai Province. The bird movement data are from 29 bar-headed geese (Anser indicus) from Qinghai Lake. Fourteen of them were captured on March 25-31, 2007, and the others were captured on March 28 - April 3, 2008. Each bird was weighed, measured and equipped with a 45g solar-powered portable transmitter terminal (PTT:9 North Star Science and Technology, LLC, Baltimore, Maryland USA) and 1 Microwave Telemetry (PTT-100, Columbia, Maryland USA). Transmitter signals were received by Argos data system (CLS America Inc., Maryland, USA) and transmitter locations were estimated. Argos classified the location accuracy into seven categories: 3, 2, 1, 0 and LA, B, Z with the approximation for class 3 < 150 m, class 2 = 150-350 m, class 1 = 350- 1000 m, class 0 > 1000 m. We also bind the GPS (Global Position System) location equipment on the PTTs. We call the location data as LG.

Table 1. Relational representation of our bird migration data

Obs	animal	ppt	date	time	Latitude	longitude	K94	Speed
85	BH07_67695	67695	2008-03-02	3:27:10	29.275	88.731	LZ	32
86	BH08_67688	67688	2008-03-02	4:27:10	30.275	89.25	KG	43

Our data sets received from Western Ecological Research Center are represented by the form shown in Table 1, which consists of 66796 and 22951 location data records for the 2007 survey and 2008 survey, respectively. About 90.1% of data records in the four categories of 0-3 are with high quality, which are used in our study; the remaining LA, B, Z categories were dismissed due to high noise. We note that PTT were deployed on 14 bar-headed geese from Qinghai Lake in March 2007. Three PTTs were still active as of 1 Nov 2008, and three PTTs were lost before the birds returned to their wintering place. In addition, among the PTTs deployed on the 15 bar-headed geese from Qinghai Lake in March 2008, nine out of them are still active as of Nov 1, 2008. Most of them have arrived at the winter area by Nov 1 2008.

We also note that for the satellite transmitters are expensive, it was impossible for us to use this equipment to track all the birds. Instead, only some key species were tracked. But many water bird species are highly faithful to the sites they use throughout their annual cycle (both within and between years) [6,7]. Such fidelity can be explained as a result of various selective pressures that flavor individuals which have an intimate knowledge of their environment. For most birds from the same population, they have the similar migration routes and habitat area [21]. Thus, although the number of our data samples is limited, the reliability and credibility of our survey are high.

4 Framework of Our Bird Migration Data Mining System

We propose a data mining system to discover the habitat area and migration route efficiently. A new hierarchical clustering algorithm is developed in the system to find sub-areas with a dense location points relative to the entire area. Then the Minimum Convex Polygon Home Range of bird species is calculated. Then, association analysis is used to discover the site connectedness and migration route between the discovered habitats. Figure 2 shows a diagram and component flow of our system, which consists of four phases: preprocessing, clustering, home range calculation and sequence mining. Each phase is described in detail in the subsequent subsections.

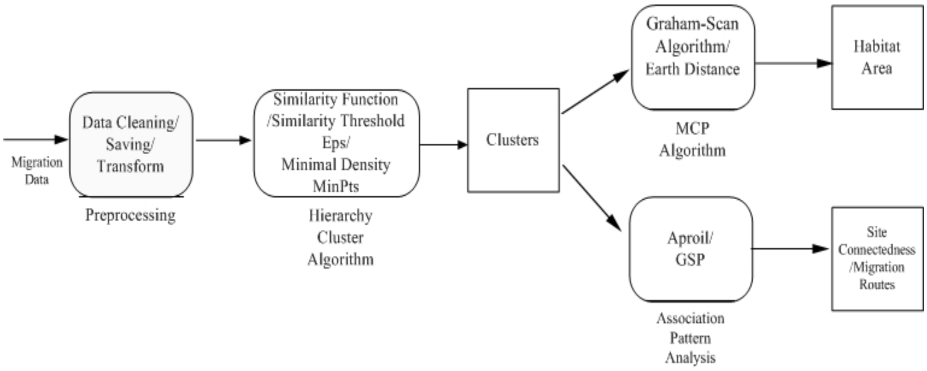

Migration Data
Data Cleaning/ Saving/ Transform
Preprocessing
Similarity Function /Similarity Threshold Eps/ Minimal Density MinPts
Hierarchy Cluster Algorithm
Clusters
Graham-Scan Algorithm/ Earth Distance
MCP Algorithm
Habitat Area
Aproil/ GSP
Association Pattern Analysis
Site Connectedness /Migration Routes

Fig. 1. System Framework of Our Bird Migration Data Mining System

4.1 Preprocessing

The raw data are downloaded from the USGS website. We focus on dynamic attributes such as latitude, longitude, time and speed. Outlier records are removed, and missing values are estimated and considered. The processed data are then stored at a relation database for further use.

4.2 Clustering Phase: Hierarchical Clustering and Spatial-Tree Building Based on DBSCAN

The objective of this phase is to mine interesting clusters from the preprocessed data set. As there are many choice of clustering algorithms, we require a clustering algorithm to satisfy the following criteria: (i) The algorithm should not require manual setting on the number of clusters. It is unreasonable to determine these parameters manually in advance. (ii) Since we only want to find important habitat area, the algorithm should filter out those with lower density. (iii) The location data are very large, the algorithm should be capable of handing a large data set within reasonable time and space constraints.

The DBSCAN [10] algorithm is a good choice as it meets all of these requirements. It does not need to input the number of clusters as a predefined parameter. According to the density-based definition, the density associated with a point is obtained by counting the number of points in a region of a specified radius, Eps, around the point. Points whose densities are above a specified threshold, MinPts, are classified as core points, while noise points are defined as non-core. Core points within the same radius of Eps to each other are merged together. Non-core and non-noise points, which are called border points, are assigned to the nearest core points. Those core points build the skeleton of a cluster. The algorithm makes use of the spatial index structure (R*-tree) to locate points within the Eps distance from the core points of the clusters. The time complexity of DBSCAN is O(N*logN). It is accepted in our application.

Biologists need to evaluate the “core areas”, and then to identify the actual areas that are used within bird home ranges. The core area usually defined as areas concentrated by individual at each wetland. For example, the fostering place would be the core region, but the foraging area would be the out-of-core region. Motivated by this requirement, we introduce a Hierarchical DBSCAN (HDBSCAN) clustering approach, which can build up a Spatial-Tree encoding every cluster node like a Huffman tree code in a top-down manner.

The pseudo code of the HDBSCAN algorithm is shown in Fig 2. It adopts a Breath First Search-like strategy that clusters the data sets by using DBSCAN. Inputs are parameter Eps and Minpts for DBSCAN, together with bird migration data and predefined tree height. By one "first in, first out" queue “Q”, spatial tree in Fig 2 is the output results. For each node in the same level of the tree, two pointers “front” and “last” point out their level (line3), and those nodes share the same DBSCAN parameter. At first, the DBSCAN are applied on those nodes (line 7). The clustering results are then put into “Q” (lines 11-15). If the depth of tree reaches the predefined tree height (lines 8-9), the hierarchical algorithm returns. Meanwhile, the id of a cluster is joined by its own cluster label and its father id as the tree grows. For instance in the

```
  Input: Location data: LD, Parameter: Eps and Minpts, S-Tree: Height
  Output: LD with cluster lable and Spatial_Tree was built
DBSCAN_ OBJECT Root=Joint(LD,Eps,Minpts); // root node of Tree
ENQUEUE(Q, Root) ;          // push DBSCAN object into Queue
front:=0, last:=0, level=0;
while(Queue<>empty and front<=last) DO
  DBSCAN_ OBJECT node= DEQueue(Q); // Pull data from Queue
   front++;               //
   Data_OBJECT  Childern    =DBSCAN.getCluster(node); //Call DBSCAN
   if(level > Height)
      break;

    For i FROM 1 TO Childern.size  DO
      Data child=Childern.get(i);
      DBSCAN_ OBJECT Root=Joint(child,Eps,Minpts);
      ENQUEUE(Q,DBSCAN_ OBJECT) ;
    end For

    if(front>last) // members in one level have been searched
       last= Q.size()+front-1;
       level ++;
    end if
end while
```

Fig. 2. Pseudo code of the HDBSCAN algorithm

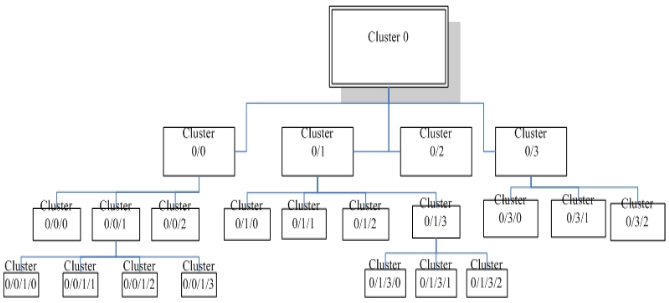


Fig. 3. An Spatial Tree Example: A Huffman coding-like Structure built by HDBSCAN

Fig 3, the left most leave node in the tree is encoded by his father id "0/0/1" and its own id "/0". Thus, its id is "0/0/1/0". By this Huffman encoding-like method, the cluster id is unique and the spatial tree is easy to manage.

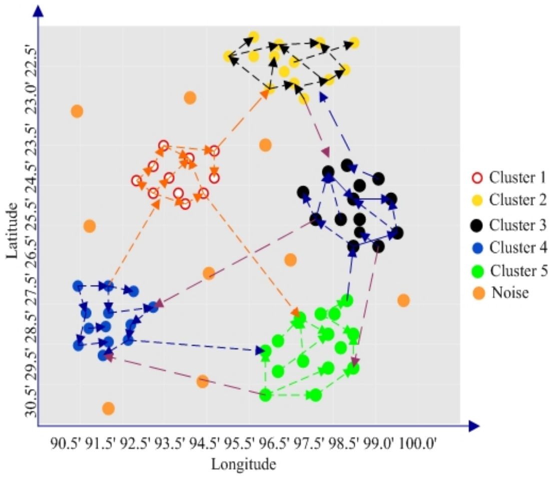
4.3 Habitat Home Range Calculation Phase

In this phase, we use the idea of MCP (Minimum Convex Polygon) to circle the clusters and spherical geometry to obtain bird species' home range. There are two algorithms that compute the convex hull of a set of n points. Graham's scan runs in O(nlgn) time complexity, and the Jarvis's march runs in O(nh) time complexity, where h is the number of vertices of the convex hull. In our work, points with maximum or minimum latitude were found at first hand, and then we utilize Graham-Scan to compute the MCP. The run time is limited to O(n). A much more technical description of this approach can be referred to [23]. A closed geometric figure on the surface of a sphere is formed by the arcs of greater circles. The spherical polygon is a generalization of the spherical triangle [24]. If Φ is the sum of the radian angles of a spherical polygon on a sphere of earth radius R, then the area is:

$$s = [\Phi - (n-2)\pi] * R^2 \qquad (1)$$

4.4 Phase for Association Analysis

In this phase, association analysis is explored to discover site connectedness and bird migration routes. As illustrated in the Fig.4, points scattered around map are bird location sites, and their color stands for the discovered clusters labeled from the clustering phase. An arrow points out a bird migration route, which is considered as the pattern in the domain of data mining. Mining those spatial-temporal relationships between discovered habitats would be important for understanding how the different biological habitat elements interact with each other.



Bird ID	Visit Time	Visit Habitat
A	T1	Cluster 1
A	T2	Cluster 2
B	T1	Cluster 1
B	T2	Cluster 3
B	T4	Cluster {5,2}
C	T3	Cluster 2
C	T4	Cluster {5,3,2}
…	…	…
N	T_n	Cluster {1,2,5}

Fig. 4. Bird migration routes between clusters are converted into records in the right table

Biologists are interested in two types of spatial-temporal association patterns that involve sequences of events extracted from the clustered areas:

- **Non-sequential pattern**- relationships among the habitats for different birds, ignoring the temporal properties of the data. It can reveal the site connectedness.
- **Sequential pattern-** temporal relationships among the habitats for different birds, which are associated with migration routes.

One way to generate associative patterns from the migration data is to transform the spatial-temporal datasets in the Fig 4 into a set of transactions as in the Table 2. The main advantage of such approach is that we can use many of the existing algorithms to discover the association patterns that exist in the data. Different cluster areas that form the movement patterns can be recorded as the items for a bird transaction.

Table 2. Transforming the migration data into market-basket type transactions

Bird ID	**T1**	**T2**	**T3**	**T4**	**T5**	
A	Cluster 1	Cluster 2	θ	Cluster {2,1}	Cluster{2,3}	….
B	Cluster 1	Cluster{3,}	θ	Cluster{5,2}	Cluster {7,9}	….
C	θ	θ	Cluster{2}	Cluster{5,3,2}	Cluster {5,10}	….
….	….	….	….	….	….	….
N	Cluster{1,2,5}	Cluster {3,6}	θ	Cluster 10	Cluster{5,10}	….

Non-sequential associations among events only concern with the spatial cluster areas, irrespective to the timing information. The abstracted events can be transformed into a transaction format. Such representation allows us to apply the existing association rule mining algorithms. In this paper, we make use of the pioneering algorithm Apriori [15] to extract the association patterns. The following three interestingness measures are suggested to evaluate the association patterns such as one like: *cluster area A → cluster area B*.

$$\text{Support}=P(A,B) \tag{2}$$

$$\text{Confidence}=P(A,B)/P(A) \tag{3}$$

$$\text{Lift}=p(B|A)/P(B) \tag{4}$$

The support of a rule A → B is the probability that a transaction contains the code {A, B}. The confidence value of the rule denotes the conditional probability of {B} given {A}. Lift is computed to judge the correlation or the dependence between {A} and {B}. The association rule can be ranked based on an individual interestingness measure or their combinations.

If temporal information is incorporated, we can derive sequential associations among the events (cluster areas) using the existing sequential pattern discovery algorithms, such as GSP [17]. We choose to use the GSP algorithm, which was initially proposed by Agrawal et al. for finding frequent sequential patterns in the market-basket data. In the GSP approach, a sequence is represented as an ordered list of itemsets, s = <s1, s2, …, sn>. Each element sn of the sequence is subject to three timing constraints: window-size (i.e. maximum time interval among all items in the element), min-gap (i.e. minimum time difference among successive elements) and max-gap (maximum time difference among successive elements). In our paper, we have set the window-size to be 1 day, min-gap to be 0 and the max-gap to be 2 days. The above interestingness measures for non-sequential pattern need to be changed accordingly so as to measure sequential patterns. For instance, given one candidate sequence: *cluster area **A*** → *cluster area **B*** → *cluster area **C*** → *cluster area **D***. the confidence and lift are computed as follows:

$$\text{Confidence=P(A->B->C->D)/P(A->B->C)} \quad (5)$$

$$\text{Lift= P(A->B->C->D)/ \{P(A->B->C)*P(D)\}} \quad (6)$$

5 Experiment Results

We conducted many experiments to evaluate our system on the data sets of bar-headed goose (See details in the section 3). In this section, we first report an efficiency result of our HDBSCAN algorithm, we then give an interpretation on the results of the habitats discovered by using our HDBSCAN algorithm, a new hierarchical clustering approach. Then, we analyze the associative pattern to reveal the bar-headed goose migration site connectedness and routes. Finally, as a part of discussion, we present some implication advice for other research topics as well. Meanwhile, we combine our system with Goolge Map for a visualization of the distribution of the habitats and migration routes.

5.1 Spatial Distribution of Bar-Headed Goose

The Spatial-Tree of 2007 is built on the annual migration data from 2007-03 to 2008-03. In Fig. 6, the left panel is the Spatial-Tree and the right panel is the spatial distribution associated with certain nodes. The convex home range is depicted with polygons with different colors and the description is presented when the user clicks the marker with certain index. Due to page limitation, we only present the first node member associated with bar-headed goose over wintering, post-breeding, and stop over sites in the Fig 6 and details description in Table 3. From Fig 6, we can clearly find the breeding area Qinghai Lake with index 3, post breeding area Zhalin-Eling Lake with index 4. The maximum one is the wintering area in Tibet river valley with index 6 covering 9254 Km^2. It is interesting to note that one species (No. BH07_67693) moved to cluster with index 1 within Mongolia rather than stay in Qinghai Lake for breeding. The average range of the habitat area is 29045.38 Km^2.

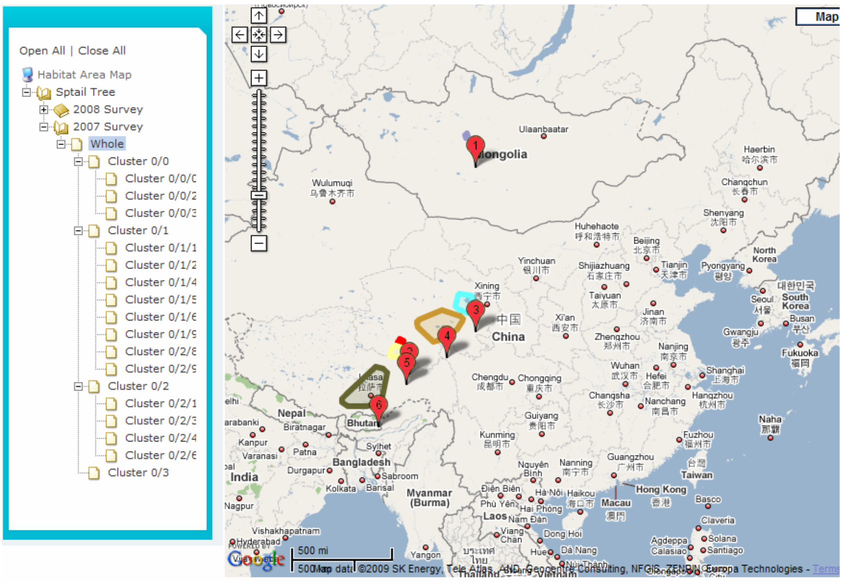


Fig. 6. Overview of 2007 bar-headed goose Spatial-Tree

Table 3. A part of results about discovered Migration habitat for bar-headed goose from Qinghai Lake, China

Cluster ID	Location Num	Home Range (Km^2)	Habitat Area Center		Bird Migration Time		Birds Num	Geography Description
			Longitude	Latitude	End	Begin		
0/0/	3368	16583.45	99.8082	36.9504	2007-10-24	2007-3-25	14	Qinghai Lake
0/1/	5394	62698.99	97.2455	35.0769	2007-12-14	2007-6-9	11	Zhalin-Eling Lake
0/3/	359	8206.649	93.4613	32.9937	2007-12-13	2007-10-11	7	YangZhiRiver YuanTou
0/2/	9069	85172.48	90.9785	29.6358	2008-2-25	2007-10-21	8	Tibet river valley
0/5/	56	361.828	99.865	47.9858	2007-6-6	2007-5-7	1	Mongolia
0/4/	34	1248.863	93.8064	33.77	2007-10-28	2007-10-13	3	TuoTuo River area

5.2 Site Connectedness of Bar- Headed Goose

As mentioned above, we transform the bird migration pattern into transaction database from in different levels of Spatial-Tree, separately. A part of association rules results from the Apriori shows some interesting patterns, where the CID means the cluster area id. Those association rules can effective evaluate the site connectedness. Those association rules can effective evaluate the site connectedness. For example, if we observe one associate rule {CID(0/0/1/) and CID(0/0/0/) and CID(0/0/3/) } -> { CID(0/1/4/) } with minimum support 21.4%. This can reveal the high site connectedness of stop area around the Qinghai Lake and Chaka salt Lake area. Thus, Qinghai lake Reserve is situated at the optimal location for storks preparing for the autumnal migration toward winter sites.

5.3 Migration Routes of Bar- Headed Goose

As described in the migration route mining in Section 4.3, bird migration between habitats could be regard as sequence. The discovered frequent sequence with higher confidence and lift would investigate a few interesting biological phenomena. A part of results are illustrated in Table 4. visualizing those sequences would help ornithologist to understand. From our observation, it is clearly that bar-headed goose departed the breeding place {ClD(0/0/)}in Qinghai Lake and then arrived at the post breeding area in Zhalin-Eling Lake area or Huangheyuan wetland {CID(0/1/)} and stayed there for about two months before heading to the south. Then they follow cluster 3 {CID(0/3/) } which is served as the stopover area, and finally arrive at the winter area {CID(0/2/)}. The movement of bar-headed goose depicted in this study conforms to the Central Asian Flyway [22]. There are eight birds migrated to Tibet river valley, and stay there from 2007-10-21 to 2008-02-25, with a total of 127 days over winter rather than fly to north-eastern India and Bangladesh. Mean fall migration duration was about 7 days. The migration distance is (1500 km (311km+382km+758km).

Table 4. Part of Sequential results in level 2 of 2007 Spatial-Tree, Minimum Support is 20% and Minimum Confidence is 30%

Rules ordered by support	1. [CID (0/0/1/)-> CID (0/0/0/)] (support=50%) 2. [CID (0/0/0/)-> CID (0/0/1)] (support=35.7%) 3. [CID ([0/1/1/)-> CID (0/2/1//)->CID (0/2/7/)-> CID (0/2/1/)] (support=21.4%)
Rules ordered by Confidence	1. [CID(0/1/1/)-> CID 0/2/1/-> CID 0/2/7/-> CID 0/2/1/] (confidence=100%) 2. [CID (0/0/0/)-> CID (0/0/0/)->CID (0/0/1/)-> CID (0/0/0/)-> CID (0/1/3/)] (confidence=100%) 3. [CID (0/0/1/)-> CID (0/1/6/) CID (0/1/3/) (confidence=75%) 4. [CID (0/0/1/)-> CID (0/0/0/)-> CID (0/1/4/)] (confidence=42%)
Rules ordered by Lift	1. CID (0/0/1/)-> CID (0/0/0/)->CID (0/0/1/)-> CID (0/3/0/) (lift=33.4%) 2. CID (0/1/1/)-> CID (0/2/1/)->CID (0/2/7/)-> CID (0/2/1/) (lift=25.4%)

5.4 Discussion and Implications for Habitat Conservation and Avian Influenza

Our cluster based approach for discovering the bar headed goose approximately depicts the geographical distribution of this species of wild water bird. Both of the cluster results in 2007 and in 2008 match greatly, which indicates that some certain habitats, such as the Qinghai Lake, DaLing Lake and the Tibet river valley, are of vital importance for some species. What is more, the discovered migration routes are critical for finding an adequate compromise between habitat protection and economic development in the regions along their migration routes. Wide areas of MCP prove that it is necessary to build a broad network to cover the different core region areas. The clustering results displayed in the GIS pave the way for human beings to construct a systematic nature reserve in future. In addition, scientists would like to do much more research, such as virus, plant and environment quality survey, to discover the way of highly pathogenic avian influenza disperse in the wild bird species' MCP.

6 Conclusion and Future Work

The satellite tracking has been used successfully to record the migration routes and stopover sites of a number of birds. Such information allows the development of a future plan for protecting the breeding and stopover sites. The proposal of our computational ideas and methods is motivated by the long-time lack of an efficient data analyzing approach which actually can help researchers to do this work systematically.

In this paper, we have suggested to explore the field by using the location data information as a supplement data mining process which can provide an alternative approach for traditional bird telemetry data analysis: visual observation from the location points. In order to discover the core range of the birds, a new clustering strategy has been introduced. This clustering strategy can effectively manage the different cluster areas and can discover the core areas in some larger habitat. Using association rule analysis, site connectedness of habitat and the autumnal migration routes for the bar-headed goose were investigated. Clustering and association rule mining do provide an effective assistance for biologists to discover new habitats and migration routes.

In the future, we plan to extend our current work to address several unresolved issues. Specifically, we intend to use Hidden Makov Model [25] to predict the bird movements. Also, we would like to compare the cluster area spatial distribution between different years, in an aim to discover the habitat changing trend of bird species. Finally, we intend to extend our analysis to other species in the Qinghai Lake to identify the cross habitat for different species.

Acknowledgments. This work was supported by Special Project of Informatization of Chinese Academy of Sciences in" the Eleventh Five-Year Plan", No.INFO-115-D02. The authors would like to thank the researchers in Western Ecological Research Center, Patuxent Wildlife Research Center and United States Geological Survey for sharing the data sets. Special Thanks to Professor Ying Liu at Graduate University of Chinese Academy of Sciences for her constructive suggestions in building our system.

References

1. Liu, J., et al.: Highly pathogenic H5N1 influenza virus infection in migratory birds. Science 309, 1206 (2005)
2. Li, Z.W.D., Mundkur, T.: Numbers and distribution of waterbirds and wetlands in the Asia-Pacific region: results of the Asian Waterbird Census 2002–2004. Wetlands International, Kuala Lumpur (2007)
3. Worton, B.J.: kernel methods for estimating the utilization distribution in home-range studies. Ecology 70, 164–168 (1989)
4. Kanai, Y., et al.: Discovery of breeding grounds of a Siberian Crane Grus leucogeranus flock that winter sin Iran, via satellite telemetry. Bird Conservation International 12, 327–333 (2002)
5. Mathevet, R., Tamisier, A.: Creation of a nature reserve, its effects on hunting management and waterfowl distribution in the Camargue (southern France). Biodiv. Conserv. 11, 509–519 (2002)

6. Shimazaki, H., et al.: Migration routes and important stopover sites of endangered oriental white storks (Ciconia boyciana) as revealed by satellite tracking
7. Ball, G.H., Hall, D.J.: ISODATA: a novel method of data analysis and pattern classification. Technical Report of Stanford Research Institute, Menlo Park, CA, Stanford Research Institute, 66 (1965)
8. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, pp. 226–231 (1996)
9. Ester, M., Kriegel, H., Sander, J., Xu, X.: Incremental Clustering for Mining in a Data Warehousing Environment VLDB (1998)
10. Ng, R.T., Han, J.: Efficient and Effective Clustering Methods for Spatial Data Mining. In: Proc. 20th Int. Conf. on Very Large Data Bases, Santiago, Chile, pp. 144–155 (1994)
11. Ertöz, L., Steinbach, M., Kumar, V.: Finding topics in collections of documents: A shared nearest neighbor approach. In: Proceedings of Text Mine 2001, First SIAM International Conference on Data Mining, Chicago, IL,USA (2001)
12. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, Inc., New York (1990)
13. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. of the 20th VLDB Conference (1994)
14. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. International Journal of Data Mining and Knowledge Discovery 8(1), 53–87 (2004)
15. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the 11th International Conference on Data Engineering (ICDE 1995) Taipei, Taipei, Taiwan, pp. 3–14 (1995)
16. Zaki, M.J.: Efficient Enumeration of Frequent Sequences. In: 7th International Conference on Information and Knowledge Management, Washington DC, November 1998, pp. 68–75 (1998)
17. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases. In: Egenhofer, M.J., Herring, J.R. (eds.) SSD 1995. LNCS, vol. 951, pp. 47–66. Springer, Heidelberg (1995)
18. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H.: Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings of the 2001 International Conference on Data Engineering (ICDE 2001), pp. 214–224 (2001)
19. Muzaffar, S.B., Johny, T.: Seasonal movements and migration of Pallas's Gulls Larus ichthyaetus from Qinghai Lake, China. Forktail 24, 100–107 (2008)
20. Miyabayashi, Y., Mundkur, T.: Atlas of key sites for Anatidae in the East Asian Flyway. Wetlands International—AsiaPacific, Tokyo: Japan, and Kuala Lumpur, Malaysia (1999), `http://www.jawgp.org/anet/aaa1999/aaaendx.htm` (accessed March 11, 2008)
21. Shan, G.H., JunYu, L., QiYing, L.: Introduction to ACM international Collegiate Programming Contest, 2nd edn., pp. 100–102 (in Chinese)
22. Weisstein, E.W.: Spherical Polygon. From MathWorld–A Wolfram Web Resource, `http://mathworld.wolfram.com/SphericalPolygon.html`
23. Daniel Sheldon, M.A.: Saleh Elmohamed, Dexter Kozen. In: Collective Inference on Markov Models for Model Appendix: Springer-Author Discount