Knowl Inf Syst
DOI 10.1007/s10115-010-0308-x

REGULAR PAPER

Exploring the wild birds' migration data for the disease spread study of H5N1: a clustering and association approach

Mingjie Tang · Yuanchun Zhou · Jinyan Li · Weihang Wang · Peng Cui · Yuanseng Hou · Ze Luo · Jianhui Li · Fuming Lei · Baoping Yan

Received: 28 October 2009 / Revised: 24 April 2010 / Accepted: 24 May 2010

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 in many decision-making processes such as for conservation site construction and avian influenza control. The raw data of the habitat areas and the migration routes are usually in large scale and with high complexity when they are determined by high-tech GPS satellite telemetry. 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 localizations, and the notion of association rules for the discovery of migration routes from the scattered location points in the GIS. One of our clustering results is a tree structure, specially called spatial-tree, which is an illusive map depicting the breeding and wintering home range of bar-headed geese. A related result to this observation is an association pattern that reveals a high possibility that bar-headed geese's potential autumn migration routes are likely between the breeding sites in the Qinghai Lake, China and the wintering sites in Tibet river valley. Given the susceptibility of geese to spread H5N1, and on the basis of the chronology and the rates of the bar-headed geese migration movements, we can conjecture that

Parts of this paper appeared in the Proceedings of the 2009 ADMA Conference [37].

M. Tang · Y. Zhou (✉) · W. Wang · Z. Luo · Jianhui Li · B. Yan
Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
e-mail: zyc@cnic.cn

P. Cui · F. Lei
Institute of Zoology, Chinese Academy of Sciences, Beijing, China

M. Tang · W. Wang · P. Cui
Graduate University of Chinese Academy of Sciences, Beijing, China

Jinyan Li
School of Computer Engineering, Nanyang Technological University, Singapore, Singapore

Y. Hou
Bureau of Qinghai Lake National Nature Reserve, Qinghai, China

bar-headed geese play an important role in the spread of the H5N1 virus at a regional scale in Qinghai-Tibetan Plateau.

Keywords Clustering · Sequence mining · Bird migration · Habitat · Route · H5N1 · Qinghai Lake

1 Introduction

The Asian outbreak of highly pathogenic avian influenza H5N1 disease in poultry in 2003, 2004 and 2009 was unprecedented in its geographical extent, and its transmission to human beings showed an ominous sign of life-threatening infection [25]. 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 bird's habitats, 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 the possibility of H5N1 is spilling over from the poultry sector into some wild bird species [10,25,40].

The spatial data analysis on the species' 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, CA, USA) ([6,17, 27]). However, there has been lack of a persuasive way to identify the stop areas 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 habitats* 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). 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 [27].
- *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 [18]. 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 [27].

Our computational approaches to addressing 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, meanwhile identifying 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 the findings. With this visualization tool, all of our results can be embedded into the Google Map[1] (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 illusive and have been confirmed to biological discovery in [19]. 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, (iii) bird migration routes are rigorously studied by sequence mining algorithms and (iv) correlation between the bird migration data and H5N1 out-breaking locations is deeply examined.

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 Sect. 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 Sect. 5. In Sect. 6, biological significance of our computational results is discussed. Finally, 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 [27,29]. Traditionally, most biologists have to count those location plots in a certain area and then utilize kernel model to calculate the home range of bird species [6,27,42]. Until recently, Hiroto et al. [18] 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 using the ISODATA algorithm [4]. 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'

1 http://maps.google.com/.

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 stand-along tool to get insight into the distribution of the data points in a data set for example sequence stream data sets [48] or can be used as a data preprocessing step for other types of data analysis as text or web mining [20]. Various techniques have been explored for clustering spatial data sets. For instance, an improved k-medoid method, called *CLARANS* [31] was proposed in the last decade. *SNN* [12] was also developed to cluster the earth science data. *DBSCAN* [13] and *IncrDBSCAN* [14] have been proposed to process the spatial data sets as well. Later, scientists Ankerst et al. [3] proposed *OPTICS* to find the suitable parameter *Eps* and *Minpts* in the *DBSCAN*.

Meanwhile, several hierarchical clustering approaches have been long investigated, including the agglomerative approach (eg. *AGNES*) and the divisive approach (eg. *DIANA*). Agglomerative methods (eg. *AGNES*) starts with as many clusters and builds a tree by successive join of the two nearest clusters. On the other hand, divisive method (eg. *DIANA*) starts with one big cluster and successive splits of the largest cluster by the most dissimilar variable. *AGNES* and *DIANA* are fully described in Chaps. 5 and 6 of [21]. *BIRCH* [46] uses a hierarchical data structure called CF-tree for partitioning the incoming data points in an incremental and dynamic way. *BIRCH* is order-sensitive as it may generate different clusters for different orders of the same input data. *CURE* [15] represents each cluster by a certain number of points that are generated by selecting well-scattered points and then shrinking them toward the cluster centroid by a specified fraction. Recently, [5] used Markov Random filed (*MRF*) to deal with multi-modality clustering in a hierarchy cluster way. In this paper, our new idea is to combine *DBSCAN* with a hierarchical clustering approach (divisive approach) to find the birds habitats with different levels of densities.

2.3 Association analysis

Association rules mining and sequence mining are pioneer research topics in data mining [22], and they are still attracting lots of attentions. The classic association rules mining algorithms include Apriori [1] and FP-tree [16]. GSP [2] was the first approach to the discovery of frequent sequence patterns. Zaki then propose the SPADE algorithm [43] to find frequent sequences at a faster speed. The PSP (Prefix Tree For Sequential Patterns) approach [32] 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. Pei et al. [33] proposed methods to discover closed frequent-patterns to reduce the number of patterns. Later, Dong et al. [11] focused on the mining of one type of comparative pattern called emerging patterns. 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 is located in the middle of Qinghai Province, and it is the largest salt

Table 1 Relational representation of bird migration data

Obs	Animal	PPT	Date	Time	Latitude	Longitude	K94	Speed (km/h)
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

Table 2 Information of satellite-tracking birds

Animal	Number	Active time begin	Active time end	Location number
Bar-headed goose	29	2007-03-21	2009-10-21	783240
Ruddy Shelduck	20	2007-03-21	2009-02-01	179302
Great black headed gull	10	2007-06-21	2008-06-07	37242

lake in China with an area of 525 km^2. 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., MD, 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 equipments on the PTTs. We call the location data as LG.

Our data sets received from Western Ecological Research Center are represented by the form shown in Table 1. For example, one location data {Obs:85; animal: BH07_67695; PPT:67695; date: 2008-03-02; time: 3:27:10; latitude: 29.275; longitude: 88.731; K94:LZ; speed: 32 km/h} stands for one bar-headed goose BH07_67695 carry the portable transmitter terminal(PPT 67695) appearing in the location (latitude: 29.275; longitude: 88.731) in 2008-03-02 3:27:10 (States Eastern Standard Time) with flying speed 32 km/h. The precision level of this location is LZ.

We also note that 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) [6,29]. 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 [24]. Thus, although the number of our data samples is limited, the reliability and credibility of our survey are high. The data sets description is summarized in the Table 2. Until 2009-10-25, bird migration number has nearly reached to one million (999784). After filtering out the duplicate location, we choose 116796 and 72951 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 November 1, 2008, and three PTTs were lost before the birds returned to their wintering place. In addition, among

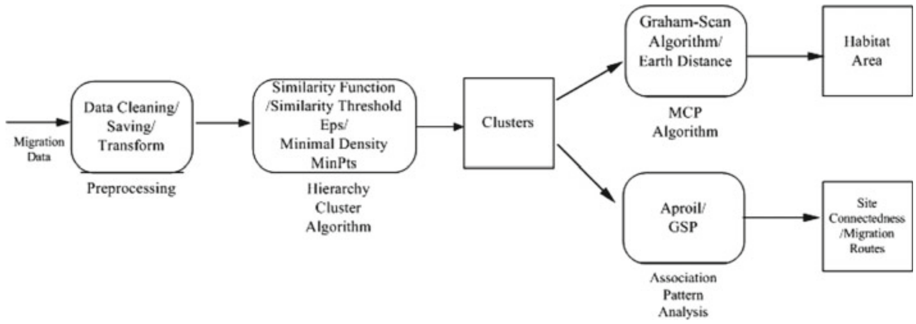


Fig. 1 Framework of our bird migration data mining system

the PTTs deployed on the 15 bar-headed geese from Qinghai Lake in March 2008, nine out of them are still active as of November 1, 2008. Most of them have arrived at the winter area by November 1, 2008.

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 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. Our system consists of four phases: preprocessing, clustering, home range calculation and sequence mining. Each phase is described in detail in the subsequent subsections. Figure 1 shows a diagram describing the component flow of our system.

4.1 Preprocessing

The raw data are downloaded from the United States Geological Survey website. We focus on dynamic attributes such as latitude, longitude, time and speed. Outlier records are removed, and missing values are estimated and considered. As the data form in the Table 1, sometimes either time or location information would be lost. We call this datasets as *missing value sample (mvs)*. As a result, we come to find *mvs's* time or location k-nearest neighbor, and use the neighbor's data to fill out missing value. For example, one *mvs M* with data information {animal: BH07_67695; PPT: 67695 date: 2008-03-02; time: 3:27:10; latitude: *missing*; longitude: *missing*;}, its location information are lost. We can find the *M's* k-nearest neighbor with time feature around 2008-03-02: 3:27:10 and animal feature is BH07_67695. Because 92.3% data are not *mvs*, the discovered k-nearest neighbors can use their equal latitude and longitude to fill out the missing location of *M*. In our system, after testing several *k* value for *k*-nearest neighbor value, we set the *k* as 4. Finally, the processed data are then stored at a relation database for the later 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 choices 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. Until recently, biologists still fail to identify the precise number of birds migration habitats along the Asia central migration route [29]; thus, it is unreasonable to determine these parameters manually in advance. (ii) Since we only want to find important habitat areas, 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 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.

Furthermore, biologists need to evaluate the *core areas*, and then to identify the actual areas that are used within bird home ranges. The core areas are usually defined as areas concentrated by individual at each wetland. The fostering place would be the core region, but the foraging area would be the out-of-core region. Meanwhile, perhaps one spring or autumn migration habitat (MH_1) is made of its subset core areas ($MH_{11}, MH_{12}, MH_{13}, \ldots MH_{1k}$). When people use cluster method to find big migration habitat sets ($MH_1, MH_2, MH_3 \ldots MH_n$), we also need to discover their corresponding subset. For this reason, the traditional density-based clustering method DBSCAN or OPTICS failing to adapt to the birds migration data sets. New approach should process data in real-time, being capable to discover several density migration habitats, while organize the results in a tree structure to search or manage clusters.

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 using DBSCAN. Inputs are parameters Eps and Minpts for DBSCAN, together with our bird migration data and a predefined tree height. By one "first in, first out" queue "Q", spatial-tree in Fig. 3 is the form of our 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 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 use a spherical geometry method 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(n $\times$ lgn) time

```
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

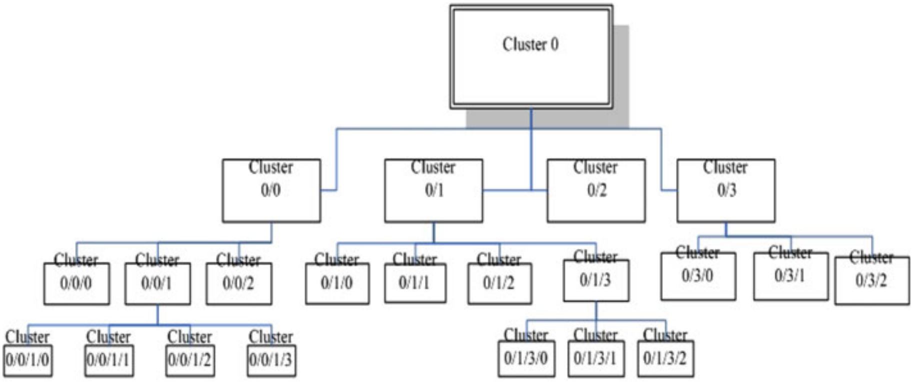
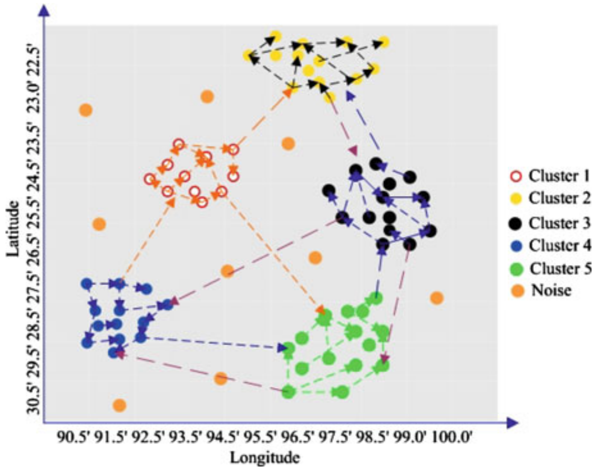


Fig. 3 A spatial-tree example: a Huffman coding-like structure built by HDBSCAN

complexity, and the Jarvis's march runs in O(n × h) 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 [35]. 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 (Weisstein Math World). If Φ is the sum of the radian angles of a spherical polygon on a sphere of earth radius R, then the area is $s = [\Phi - (n-2)\pi] * R^2$.

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,



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

Table 3 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}	….

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.

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 shown in Table 3. 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.

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 [1] to extract the association patterns. The following three interestingness measures are suggested to evaluate the association patterns like the one: *cluster area A* $\rightarrow$ *cluster area B*.

1. support $= p(A, B)$
2. *confidence* $= p(A, B)/p(A)$
3. *lift* $= p(B|A)/p(B)$

The support of a rule A $\rightarrow$ 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 [2]. We choose to use the GSP algorithm, which was initially proposed in [2], for finding frequent sequential patterns in the market-basket data. In the GSP approach, a sequence is represented as an ordered list of itemsets, $s = \langle s_1, s_2, \ldots, s_n \rangle$. Each element s_n 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* $\rightarrow$ $clusterarea\ B \rightarrow$ *cluster area C* $\rightarrow$ $clusterarea\ D$, the confidence and lift are computed as follows:

1. *confidence* $= p(A \rightarrow B \rightarrow C \rightarrow D)/p(A \rightarrow B \rightarrow C)$
2. *lift* $= p(A \rightarrow B \rightarrow C \rightarrow D)/(p(A \rightarrow B \rightarrow C) * p(D))$

5 Experiments

We conducted habitat and migration routes discovery experiment to evaluate our system on the data sets of bar-headed goose. 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 using our HDBSCAN algorithm. Then, we analyze the associative pattern to reveal the bar-headed goose migration site connectedness and routes. Finally, we present some suggestions for other research topics, then discuss our data mining results biological novelty and its association with H5N1 translating. Meanwhile, we combine our system with Google Map for a visualization of the distribution of the habitats and migration routes.

5.1 HDBSCAN: efficiency of the newly proposed hierarchical clustering algorithm

The location points in our paper are geographic coordinates (latitude and longitude), object dissimilarity are calculated by the great-circle distance as geographical distance formula (Weisstein). Our clustering input parameters *Eps* and *MinPts* were determined by Ester et al. [13], which were set as 35,000 m and 25 (and also 3,000 m and 25 in another experiment); 300 m and 25 were used for the node members in each level of the spatial-tree separately. Datasets were chosen from bar-headed goose undertake the whole pattern of spring and wither migration from 2007-03-15 to 2009-02-20. Its number is 112874 after preprocessing to filter out duplicate. Figure 5 shows the runtime of HDBSCAN by varying the depth of the spatial-trees from 4 to 1. The result illustrates that when the depth of the spatial-trees

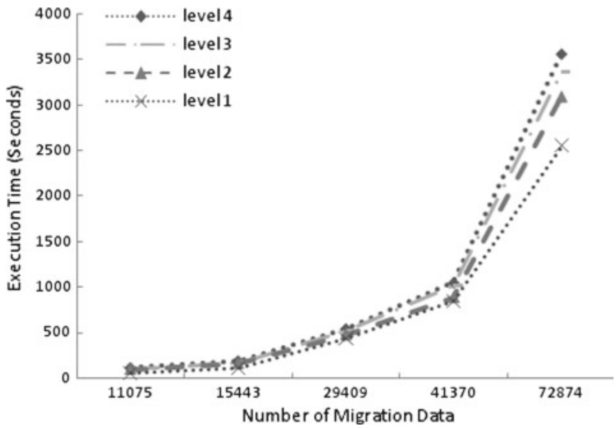


Fig. 5 Execution time: hierarchical clustering on bar-headed goose data when the number of data points varies. The number of records varied from 11076 to 72874

increases, HDBSCAN can still handle it efficiently. It is clear that our clustering algorithm exhibits a sub-linear scalability in runtime against the number of migration data points in the data base.

5.2 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. We 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 4. 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. In addition, Fig. 7 graphically displays a distribution of bar-headed goose around Qinghai Lake, China. Habitats in this area are the subset of the bigger cluster in Fig. 6 with index 0/0/. Especially, if biologists desire to understand any subset habitat in the Fig. 7, they can go to the bigger habitats' leaf node such as Fig. 8. Habitats in Fig. 8 are the subset of habitat with index 0/0/4 in the Fig. 7. Throughout the hierarchy clustering, the different level of birds' spatial distribution is visible. This process allows all cluster assignment in each level to be arranged in a tree structure. Another benefit of this method is that our MCP finds the points in the cluster convex rather than thousands of points.

5.3 Site connectedness of bar-headed goose

As mentioned earlier, we can transform the bird migration patterns into a transaction database at different levels of spatial-tree, separately. Some association rules (Table 5) detected by the Apriori algorithm are interesting, where the CID means the cluster area id. These

Fig. 6 Overview of 2007 bar-headed goose spatial-tree, the left of figure is the spatial-tree, the circles with different color in the right are discovered birds migration habitats

Table 4 Some discovered migration habitat for bar-headed goose from Qinghai Lake, China

Cluster ID	Location Num	Home range (km^2)	Habitat center		Bird migration time		Birds Num	Geography description
			Longitude	Latitude	End	Begin		
0/0/	3368	16583.4	99.8082	36.9504	2007-10-24	2007-3-25	14	Qinghai Lake
0/1/	5394	62698.9	97.2455	35.0769	2007-12-14	2007-6-9	11	Zhalin-Eling Lake
0/3/	359	8206.64	93.4613	32.9937	2007-12-13	2007-10-11	7	YangZhiRiver YuanTou
0/2/	9069	85172.4	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	1248.86	93.8064	33.77	2007-10-28	2007-10-13	3	TuoTuo River area

association rules can effectively evaluate the site connectedness. For example, if we observe the associate rule $\{CID_{0/0/1/}$ and $CID_{0/0/0/}$ and $CID_{0/0/3/}\} \rightarrow \{CID_{0/1/4/}\}$ with minimum support 21.4%. Then, three habitat areas including $CID_{0/0/1/}$, $CID_{0/0/0/}$ and $CID_{0/0/3/}$ around Qinghai lake with index 1,2,3 in Fig. 9 would directly lead to $CID_{0/1/4/}$with index 4 around Chaka salt lake area. This can reveal the high site connectedness of stop areas around the Qinghai Lake and Chaka salt Lake areas. Thus, Qinghai lake Reserve is situated at the optimal location for storks preparing for the autumnal migration toward winter sites.

5.4 Migration routes of bar-headed goose

As described in the migration route mining in Sect. 4.3, bird migration between different habitats could be regarded as a sequence. The discovered frequent sequences with high

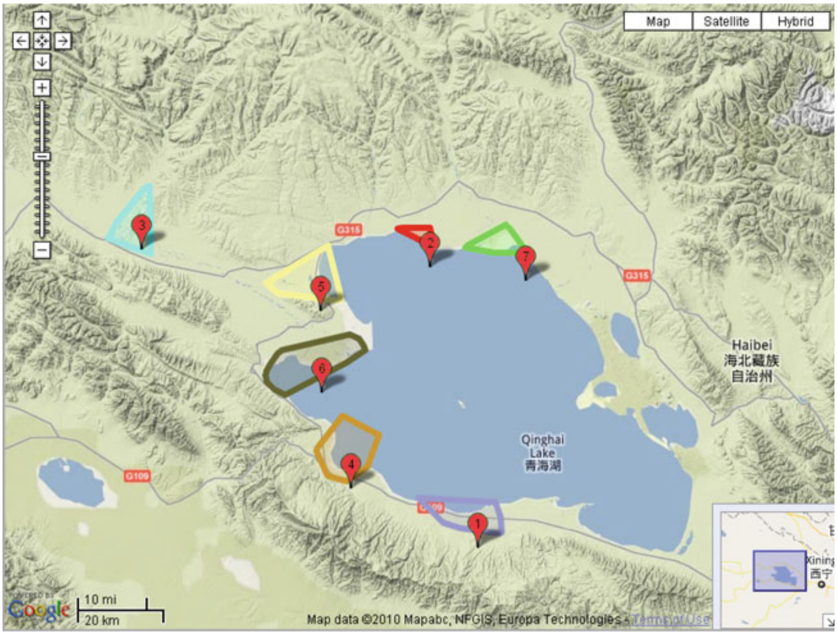


Fig. 7 Birds migration habitat spatial distribution around Qinghai Lake, China, second level of spatial-tree

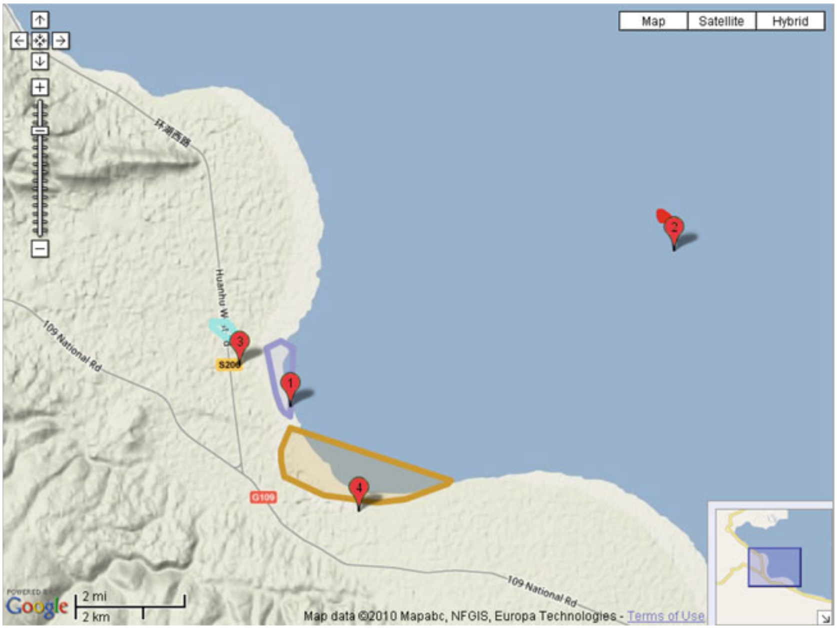


Fig. 8 Birds migration habitat spatial distribution around San Kuaisi island, Qinghai lake, China, third level of spatial-tree

Table 5 Association rules at the level 2 of the 2007 spatial-tree where the minimum support is 20% and minimum confidence is 80%

Rules ordered by support	1. $CID_{0/1/6/} \rightarrow CID_{0/0/1/}$	Support = 35.7%
	2. $CID_{0/1/3/} \rightarrow CID_{0/0/1/}$	Support = 28.6%
	3. $\{CID_{0/2/10/}$and $CID_{0/1/3/}\} \rightarrow CID_{0/0/1/}$	Support = 21.4%
	4. $\{CID_{0/0/4/}$ and $CID_{0/1/2/}$ and $CID_{0/0/1/}\} \rightarrow CID_{0/0/0/}$	Support = 21.4%
	5. $\{CID_{0/0/1/}$ and $CID_{0/0/0/}$ and $CID_{0/0/3/}\} \rightarrow CID_{0/1/4/}$	Support = 21.4%
Rules ordered by confidence	1. $CID_{0/1/4/} \rightarrow CID_{0/0/1/}$	Confidence = 100%
	2. $\{CID_{0/1/3/}$and $CID_{0/2/10/}\} \rightarrow CID_{0/0/1/}$	Confidence = 100%
	3. $\{CID_{0/0/0/}$and $CID_{0/1/2/}$and $CID_{0/0/4/}\} \rightarrow CID_{0/0/1/}$	Confidence = 100%
	4. $CID_{0/1/6/} \rightarrow CID_{0/0/1/}$	Confidence = 80%
Rules ordered by lift	1. $\{CID_{0/0/4}/$ and $CID_{0/0/0/}\} \rightarrow CID_{0/1/2/}$	Lift = 4.66
	2. $\{CID_{0/2/4/}$ and $CID_{0/0/1/}\} \rightarrow CID_{0/1/2/}$	Lift = 3.5
	3. $\{CID_{0/2/4/}$and $CID_{0/0/1/}\} \rightarrow CID_{0/1/2/}$	Lift = 3.5
	4. $\{CID_{0/1/2/}$ and $CID_{0/0/1/}$ and $CID_{0/0/4/}\} \rightarrow CID_{0/0/0/}$	Lift = 1.556

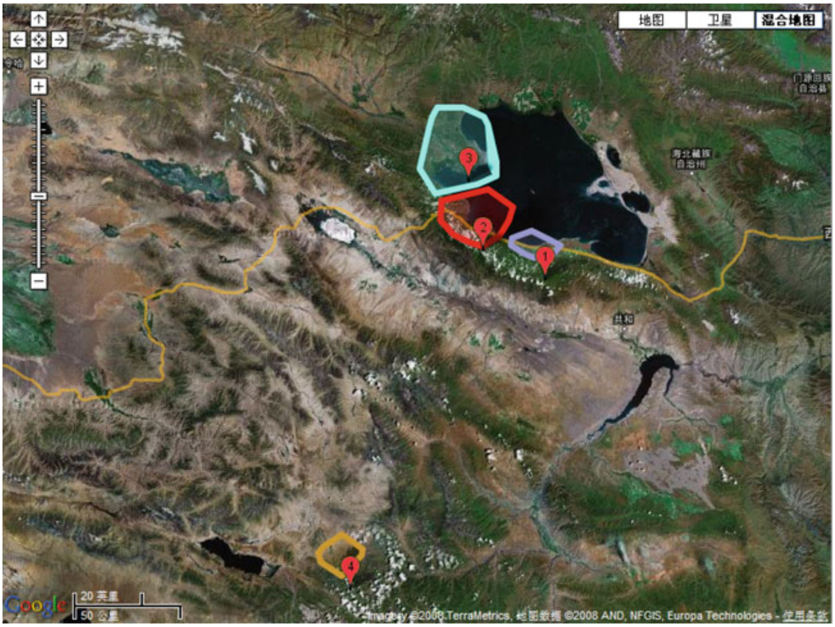


Fig. 9 Association rule from Qinghai Lake to Chaka salt lake

confidence and lift would be useful for ornithologist to investigate a few interesting biological phenomena. A part of our results are illustrated in Table 6. For example in Fig. 10, we can observe the main movements of the fourteen bar-headed geese from the Sequence $\{ClD_{0/0/} \rightarrow CID_{0/1/} \rightarrow CID_{0/3/} \rightarrow CID_{0/2/}\}$. In Fig. 10, those cluster ids are with index 1,2,3,4 separately. From our observation, it is clearly that bar-headed goose departed the breeding place $\{CID_{0/0/}\}$ 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

Table 6 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. [$CID_{0/0/1/} \rightarrow CID_{0/0/0/}$]	Support = 50%
	2. [$CID_{0/0/0/} \rightarrow CID_{0/0/1}$]	Support = 35.7%
	3. [$CID_{0/1/1/} \rightarrow CID_{0/2/1/} \rightarrow CID_{0/2/7/} \rightarrow CID_{0/2/1/}$]	Support = 21.4%
Rules ordered by confidence	1. [$CID_{0/1/1/} \rightarrow CID_{0/2/1/} \rightarrow CID_{0/2/7/} \rightarrow CID_{0/2/1/}$]	Confidence = 100%
	2. [$CID_{0/0/0/} \rightarrow CID_{0/0/0/} \rightarrow CID_{0/0/1/} \rightarrow CID_{0/0/0/}$ $\rightarrow CID_{0/1/3/}$]	Confidence = 100%
	3. [$CID_{0/0/1/} \rightarrow CID_{0/1/6/} CID_{0/1/3/}$	Confidence = 75%
	4. [$CID_{0/0/1/} \rightarrow CID_{0/0/0/} \rightarrow CID_{0/1/4/}$]	Confidence = 42%
Rules ordered by lift	1. [$CID_{0/0/1/} -> CID_{0/0/0/} \rightarrow CID_{0/0/1/} \rightarrow CID_{0/3/0/}$]	Lift = 33.4%
	2. [$CID_{0/1/1/} \rightarrow CID_{0/2/1/} \rightarrow CID_{0/2/7/} \rightarrow CID_{0/2/1/}$]	Lift = 25.4%

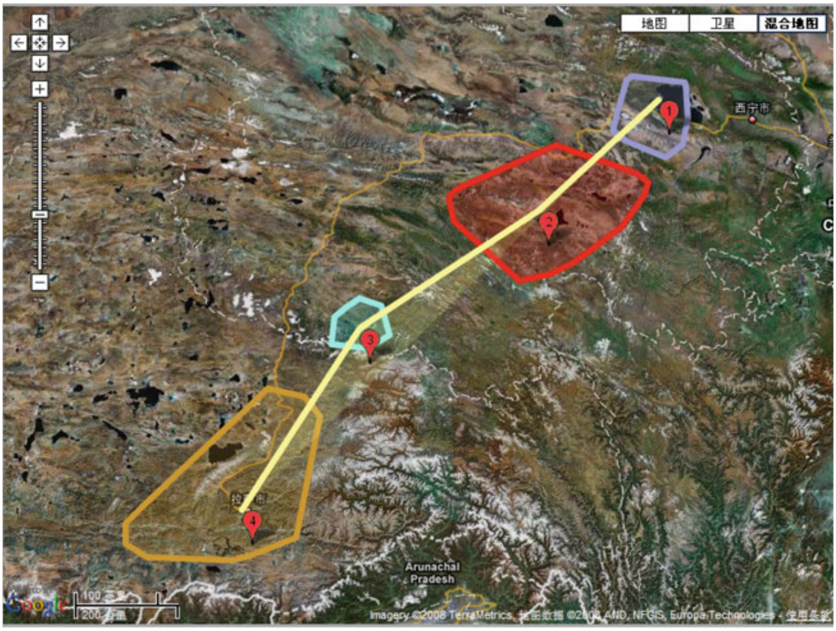


Fig. 10 Long distance migration route from Qinghai Lake to Tibet river valley in 2007

served as the stopover area, and finally arrived at the winter area $\{CID_{0/2/}\}$. The movement of bar-headed goose depicted in this study conforms to the Central Asian Flyway [29]. There are eight birds migrated to Tibet river valley, and stayed there from 2007-10-21 to 2008-02-25, with a total of 127 days over winter rather than flied to north-eastern India and Bangladesh. Mean fall migration duration was about 7 days. The migration distance was 1500 km (311 km + 382 km + 758 km).

5.5 Validation results and discussion

In the experiment, we evaluate our clustering and association rule mining results. One of the most important issues in cluster analysis is the evaluation of clustering results to find

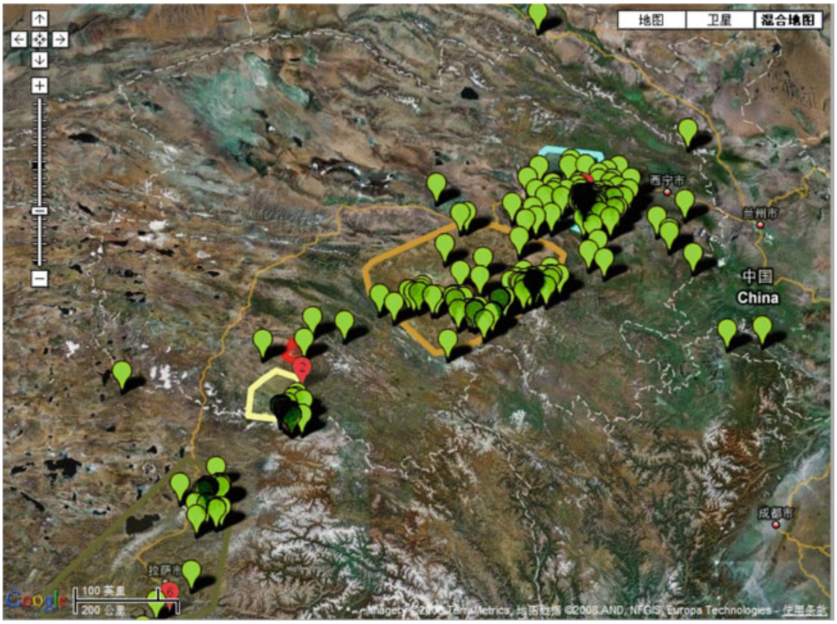


Fig. 11 Birds migration location overlapping with discovered habitats. Green points are parts of 2008 spring bar-headed goose migration location points. Circles with different color are migration habitats discovered from 2007 spring bar-headed goose migration

the partitioning that best fits the underlying data [26]. Maria though there are three major clustering validation methods: *external criteria, internal criteria and relative criteria.* In this work, our datasets is 2D, and reader is able to visually verify the validity of the results [26].

Furthermore, wild species season migration will not change greatly when the environment influence factors fail to vary dramatically [8,47]. This is called wild species migration loyalty. In case of our bird's migration dataset, one more relevant measure for the performance of the algorithm for our purpose is overlapping different year migration data sets. For example, one group of bar-headed goose {BH07_82080, BH07_82090, BH07_82081, BH07_82079, BH07_82086, BH07_82087, BH07_82091, BH07_82092, BH07_82093} would fly around same area over year. As a result, we at first use data mining model to discover bird migration habitats and routes from bar-headed 2007 migration data sets with time varied from March 25th of 2007 to March 25th of 2008. Then we use same group of bar-headed goose migration location points to map on the discovered results as in the Fig. 11, and we can count the number of location points dropping in the range of mining results. According our experiment, 89.4% 2008 migration location points precisely drop in the range of our 2007 discovered habitats, and 78.61% points are located around our discovered migration routes. Excluding some incidental migration points, as hoped, our clustering and association rule precisely reflect birds migration.

5.6 Wild birds migration routes in comparison with H5N1 outbreaks

Understanding the role which migratory water birds played in the ecology and transmission of H5N1 requires the integration of various types of data, including the habitat data, seasonal movement chronology and the routes, dates and H5N1 outbreak locations, and requires

Table 7 Migration habitats of bar-headed goose in the 2008 autumn migration and the correlated H5N1 outbreaks

Habitat ID	Number of location points	Birds' dwelling (TimeDay)	Number of birds	Name description	H5N1 once out-breaking time
0/2	23533	144	15	Qinghai Lake Area	2006-04-23
0/3	6704	110	9	ZhaLing Lake Area	2006-04
0/4	278	14	7	Stopover Area	2006-04
0/5	702	22	6	Naqu Area	NO
0/6	6983	125	6	Lasha Area	2005-08-10 2007-03-01 2008-01-25 2008-02-06 2008-03-12 2009-04-12

comprehensive analyses at both temporal and spatial levels [30]. Information about the H5N1 outbreaks was obtained from the Ministry of Agriculture of the People's Republic of China Database and OIE Database for the period between 16 February 2004 and 18 May 2009. One bar-headed goose migration route ($CID_{0/2} \rightarrow CID_{0/3} \rightarrow CID_{0/4} \rightarrow CID_{0/5} \rightarrow CID_{0/6}$, support = 21%, confidence = 100%, lift = 20%) is discovered from the bar-headed gooses migration location datasets, where all of the selected birds have undertaken the whole autumn migration from 2008-02-01 to 2009-02-01.

Our computational results and observation are illustrated in Fig. 12 and Table 6. It can be noted that the correlation between the birds' migration behavior and the timing of the H5N1 incidents in water birds regions around Qinghai-Tibet Plateau, China was very high. Nine of the H5N1 outbreaks involving water birds occurred during the winter, when the potential for interaction with poultry and the probability of direct transmission from poultry to migratory water birds were believed to be the highest. For instance, the habitats in Table 6 with cluster id (0/5, 0/6) are the most important areas for water birds to overwinter, and with a high density of population and poultry [10,24]. One of the habitats (0/6) is located at Lasha area with a bird location number 6983, and 6/21 birds prefer to stay in this area for 125 days (Table 7).

6 Discussions

Our clustering method for identifying the habitats and migration routes of the bar-headed goose is capable of depicting the geographical distribution of this species of wild water bird. Both the clustering results in 2007 and in 2008 match to each other very well, indicating that some certain habitats, such as the Qinghai Lake, DaLing Lake and the Tibet river valley, are of vital importance for their living. Moreover, the migration routes of the bar-headed goose identified in this study were part of the Central Asian flyway [28,45], in addition to those of (*Tufted duck Aythya fuligula*, *great crested grebe, Eurasian wigeon Anas penelope*, *mallard A. platyrhynchos)*, and *ruddy shelduck*, etc. The discovered migration routes are critical for finding a good 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 those related to virus, plant

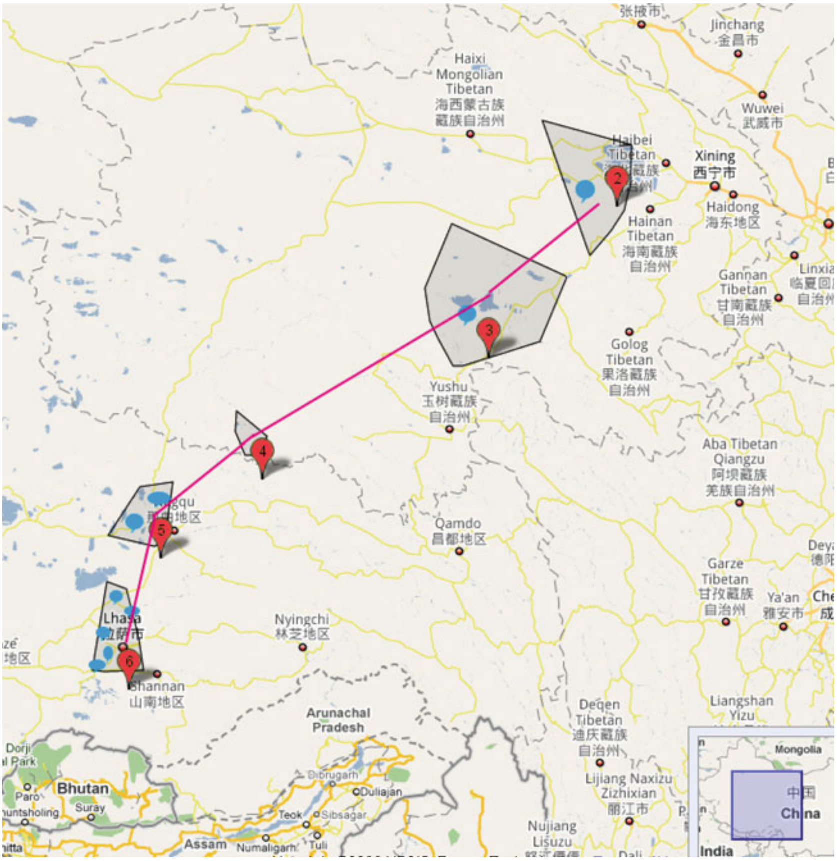


Fig. 12 An autumn migration route (0/2 → 0/3 → 0/4 → 0/5 → 0/6) of one bar-headed goose in 2008 in comparison with the H5N1 outbreak locations involving both wild and domestic birds. *Circles in black* represent habitats, *lines in red* are the migration routes, and *diamond circles in blue* are the H5N1 outbreak locations

and environment quality survey, to discover the way of highly pathogenic avian influenza disperse in the wild bird species' MCP [23].

In this work, we integrated data of migration chronology, discovered routes and habitats of satellite marked bar-headed goose, dates and locations of H5N1 incidents in Qinghai-Tibet Plateau, and analyzed spatial and temporal correlations between the bar-headed goose movements and H5N1 incidents at Qinghai-Tibet Plateau. The results revealed that the bar-headed goose play an important role in the spread of HPAI H5N1 at Qinghai-Tibet Plateau. Several studies at Qinghai Lake have also proposed that H5N1 viruses in Qinghai Lake could be spread with bird migration before [24,25,29,39]. These previous results were from both viruses' sequences analyses and satellite tracking of wild water birds. Most of their analyses were in a relative large-scale and manual analysis e.g. between countries). This study focused on an analysis at a fine regional scale within Qinghai-Tibet Plateau. Bar-headed goose in Qinghai-Tibet Plateau is a quite good model to study the role of wild migratory birds in the long distance spread of HPAI H5N1 virus.

The sites where wild water birds were infected by the H5N1 outbreaks were quite consistent with the bar-headed goose movements. We speculate that bar-headed goose were the most likely vector for the H5N1 virus' invasion into these sites. Under this conjecture, some bar-headed goose may transmit H5N1 virus asymptomatic or some bar-headed goose may transmit H5N1 virus between these sites before they die after they have contacted the disease [8,47]. Our research findings may also provide some clue toward the argument in Europe [34] that wild ducks would contract H5N1 from poultry and then translate it over great distances as they migrate, and bar-headed goose or other migratory water birds that die from H5N1 in the region would have been infected by the disease via secondary contact including exposure to fresh faces or contaminated water or vegetation. The migration routes also could be considered as that water birds are directly infected from poultry i.e., spillover, and they may be responsible for local movement of virus regionally, followed by the potential to transmit virus back to poultry i.e., spillback [36].

7 Conclusions 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 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 this study, we evaluate temporal and spatial correlation between the birds' migration routes and H5N1 outbreak location; we may draw a conclusion that bar-headed goose was an important vector of H5N1 virus in Qinghai- Tibet Plateau.

In the future, we plan to extend our to address several unresolved issues. At first, we need to discover the maximal frequent pattern as for bird migration routes by PADS [44]. Secondly, our results validation methods are simple, in the future, we would use several cluster validation algorithm [9] and field survey to test our clustering and sequence mining results. Specifically, we found that bird migration routes in small range of area usually are graph patterns rather than simple sequence, thus we intend to use graph mining to find highly correlated subgraphs (clique or quasi-clique) [38]. Finally, we intend to extend our analysis to deal with increment datasets [14] and focus on the different distance considering time attribute measure [7].

Acknowledgments 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. The authors are also very grateful to Mr. Zhang Haiting and Mr. Hu Liangling for their contribution to birds' migration datasets preprocessing, to the anonymous reviewers of our ADMA2009 and KAIS submissions for their constructive comments, which helped improve the quality of the paper significantly. The authors are also grateful to Professor Qiang Yang, Jian Pei, Xiaofeng Meng and Joao Gama for their prompt and expert handling of this article. This research was supported by National Natural Science Foundation of China (NSFC) granted by "e-Science research environment building based on networking", CAS Innovation Program (KSCX2-YW-N-063), Chinese Academy of Sciences Scientific database (No. INF105-SDB), Special Project of Informatization of Chinese Academy of Sciences in "the Eleventh Five-Year Plan" No.INFO-115-D02, "Special Project of Informatization of Chinese Academy of Sciences in "the Eleventh Five-Year Plan", Basic Databases of Joint Research Center of Chinese Academy of Sciences and Qinghai Lake National Nature Reserve under Grant No. INFO-115-C01-SDB2-02". Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and not necessarily reflect the views of the funding agencies.

References

1. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on Very Large Data Bases(VLDB'94), Santiago, Chile, pp 487–499
2. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the 11th international conference on Data Engineering (ICDE'95) Taipei, Taiwan, pp 3–14
3. Ankerst M, Breunig MM, Kriegel H-P, Sander J (1999) OPTICS: ordering points to identify the clustering structure. In: ACM SIGMOD Record, vol 28, Nr. 2, New York, NY, USA: ACM 1999, pp 49–60
4. Ball GH, Hall DJ (1965) ISODATA: a novel method of data analysis and pattern classification. In: Technical report of Stanford Research Institute, Menlo Park, CA, Stanford Research Institute, pp 66
5. Bekkerman R, Scholz M, Viswanathan K (2009) Improving clustering stability with combinatorial MRFs. In: The 15th ACM conference on knowledge discovery and data mining (SIGKDD09), pp 99–107
6. Berthold P, Terrill SB (1991) Recent advances in studies of bird migration. Ann Rev Ecol Syst 22:357–378
7. Brecheisen S, Kriegel H-P, Pfeifle M (2006) Multi-step density-based clustering. Knowl Inf Syst 9(3):284–308
8. Brown JD, Stallknecht DE, Swayne DE (2008) Experimental infection of swans and geese with highly pathogenic avian influenza virus (H5N1) of Asian lineage. Emerg Infect Dis 14:136–142
9. Chen K, Liu L (2009) "Best K": critical clustering structure in categorical datasets. knowl Inf Syst 20(1):1–33
10. Chen H, Smith GJ, Zhang SY, Qin K, Wang J, Li KS, Webster RG, Peiris JS, Guan Y (2005) Avian flu: H5N1 virus outbreak in migratory waterfowl. Nature 436:191–192
11. Dong G, Li J (2005) Mining border descriptions of emerging patterns from dataset pairs. Knowl Inf Syst 8:178–202
12. Ertöz L, Steinbach M, Kumar V (2001) Finding topics in collections of documents: A shared nearest neighbor approach. In: Proceedings of Text Mine'01, First SIAM international conference on Data Mining SDM'01), Chicago, IL, USA
13. Ester M, Kriegel H-P, Sander J, Xu X (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. Portland, OR, pp 226–231
14. Ester M, Kriegel H-P, Sander J, Wimmer M, Xu X (1998) Incremental clustering for mining in a data warehousing environment. In: Proceedings of the 24th international conference on Very Large Data Bases conference (VLDB'98), New York, USA, OR, AAAI Press, pp 226–231, 1996
15. Guha S, Rastogi R, Shim K (1998) CURE: an efficient clustering algorithm for large databases. In: Proceedings of the 1998 ACM SIGMOD international conference on Management of Data (SIGMOD'98) New York, NY, USA, pp 73–84
16. Han J, Pei J, Yin Y, Mao R (2004) Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Proc Int J Data Min Knowl Discov (81):53–87
17. Higuchi H (1991) Cooperative work on crane migration from Japan to the U.S.S.R. through Korea and China. In: Salathé T (ed). Conserving migratory birds Cambridge, International Council for Bird Preservation, pp 189–201
18. Hiroto S, Tamura M, Higuchi H (2004) Migration routes and important stopover sites of endangered oriental white storks Ciconia boyciana as revealed by satellite tracking. Mem Natl Inst Polar Res (special issue) 58:162–178
19. Kamal Rp, Tosh C, Pattnaik B, Behera P, Nagarajan S, Gounalan S, Shrivastava N, Shankar Bp, Pradhan Hk (2007) Analysis of the PB2 gene reveals that Indian H5N1 influenza virus belongs to a mixed-migratory bird sub-lineage possessing the amino acid lysine at position 627 of the PB2 protein. Arch Virol 152:1637–1644
20. Kandylas V, Upham SP, Ungar LH (2008) Finding cohesive clusters for analyzing knowledge communities. Knowl Inf Syst 17(3):335–354
21. Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. John Wiley, New York

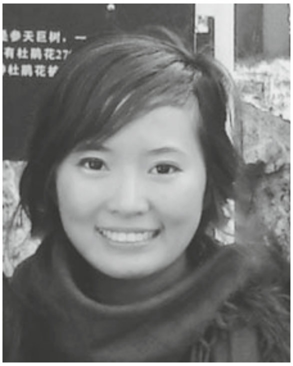
22. Koperski K, Han J (1995) Discovery of spatial association rules in geographic information databases. In: Proceedings of the 4th international symposium on large spatial databases (SSD95), pp 47–66
23. Kou Z, Li Y, Yin Z, Guo S, Wang M et al (2009) The survey of H5N1 flu virus in wild birds in 14 provinces of China from 2004 to 2007. PLoS ONE 4(9):e6926. doi:10.1371/journal.pone.0006926
24. Lei F, Tang S, Zhao D, Zhang X, Kou Z, Li Y, Zhang Z, Yin Z, Chen S, Li S, Zhang D, Yan B, Li T (2007) Characterization of H5N1 influenza viruses isolated from migratory birds in Qinghai province of China in 2006. Avian Dis 51:568–572
25. Liu J et al (2005) Highly pathogenic H5N1 influenza virus infection in migratory birds. Science 309(5738):1206
26. Maria H, Yannis B, Michalis V (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145
27. Mathevet R, Tamisier A (2002) Creation of a nature reserve, its effects on hunting management and waterfowl distribution in the Camargue southern France. Biodiv Conserv 11:509–519
28. Miyabayashi Y, Mundkur T (1999) Atlas of key sites for Anatidae in the East Asian Flyway. Tokyo: Japan, and Kuala Lumpur, Malaysia: Wetlands International—AsiaPacific.Available at http://www.jawgp.org/anet/aaa1999/aaaendx.htm. Accessed 11 March 2008
29. Muzaffar SB, Johny T (2008) Seasonal movements and migration of Pallas's Gulls Larus ichthyaetus from Qinghai Lake, China. Forktail 24(2008):100–107
30. Newman SH, Iverson SA et al (2009) Migration of whooper swans and outbreaks of highly pathogenic avian influenza H5N1 virus in Eastern Asia. PLos ONE 4(5)
31. Ng RT, Han J (1994) Efficient and effective clustering methods for spatial data mining. In: Proceedings of the 20th international conference on Very Large Data Bases (VLDB'94), Santiago, Chile, pp 144–155
32. Pei J, Han J, Mortazavi-Asl B, Pinto H (2001) Prefixspan: mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings of the 2001 International Conference on Data Engineering (ICDE 01), pp 214–224
33. Pei J, Dong G, Zou W, Han J (2004) Mining condensed frequent-pattern bases. Knowl Inf Syst 6:570–594
34. Sabirovic M, Wilesmith J et al (2006) Defra. Situation analysis—outbreaks of HPAI H5N1 virus in Europe during 2005/2006— an overview and commentary. In: International Animal Health Division, 1A Page Street, London, SW1P 4PQ, United Kingdom. Version 1, Released 30 June 2006, p 40
35. Shan G, Li J, Lin Q (2005) Introduction to ACM international collegiate programming contest, 2nd edn. pp 100–102 (in Chinese)
36. Sturm-Ramirez KM, Hulse-Post DJ, Govorkova EA, Humberd J, Seiler P et al (2005) Are ducks contributing to the endemicity of highly pathogenic H5N1 influenza virus in Asia? J Virol 79:11269–11279
37. Tang M, Zhou Y, Cui P, Wang W, Li J, Hou Y-S, Yan B (2009) Discovery of migration habitats and routes of wild bird species by clustering and association analysis. In: The 5th international conference on advanced data mining and applications. LNAI 5678, pp 288–301
38. Tang M, Wang W, Jiang Y, Zhou Y, Li J, Cui P, Liu Y, Yan B (2010) Birds bring flues? Mining frequent and high weighted cliques from birds migration networks. In: The 15th international conference on Database Systems for Advanced Applications (DASFAA2010), LNCS 5982, pp 359–370
39. Wang G, Zhan D, Li L, Lei F, Liu B, Liu D, Xiao H, Feng Y, Li J, Yang B, Yin Z, Song X, Zhu X, Cong Y, Pu J, Wang J, Liu J, Gao GF, Zhu Q (2008) H5N1 avian influenza re-emergence of Lake Qinghai: phylogenetic and antigenic analyses of the newly isolated viruses and roles of migratory birds in virus circulation. J Gen Virol 89:697–702
40. Webster RG, Peiris M, Chen H, Guan Y (2006) H5N1 outbreaks and enzootic influenza. Emerg Infect Dis 12:3–8
41. Weisstein EW (1994) "Spherical Polygon." From the Math World-A Wolfram Web Resource. http://mathworld.wolfram.com/SphericalPolygon.html
42. Worton BJ (1989) Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70:164–168
43. Zaki M (1998) Efficient enumeration of frequent sequences. In: The 7th international conference on information and knowledge management, pp 68–75, Washington DC
44. Zeng X, Pei J, Wang K, Li Jinyan (2009) PADS: a simple yet effective pattern-aware dynamic search method for fast maximal frequent pattern mining. Knowl Inf Syst 20(3):375–391
45. Zhang FY, Yang RL (1997) Bird migration research of China. China Forestry Publishing House, Beijing
46. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Rec 25(2):103–114
47. Zhou Jy, Shen Hg, Chen Hx, Tong Gz, Liao M, Yang Hc, Liu Jx (2006) Characterization of a highly pathogenic H5N1 influenza virus derived from bar-headed geese in China. J Gen Virol 87:1823–1833
48. Zhou A, Cao F, Qian W, Jin C (2008) clusters in evolving data streams over sliding windows. Knowl Inf Syst 15(2):181–214

Author Biographies

Mingjie Tang is a master candidate at Graduate University of Chinese Academy of Sciences. His research activities are focused on data mining and bioinformatics. He is currently working on sequence and graph mining, and its application on ecology.

Yuanchun Zhou is an associate professor at the Computer Network Information Center, Chinese Academy of Sciences. He got his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. His main research interests include data mining, scientific workflow and data intensive computing. He has over 30 scientific publications.

Jinyan Li received his Ph.D. degree in Computer Science from the University of Melbourne in 2001. He is an associate professor in the School of Computer Engineering, Nanyang Technological University, Singapore. His research is focused on protein structural bioinformatics, statistically important discriminative patterns, interaction subgraphs and classification methods. Jinyan has published over 100 research articles. One of his most interesting works was a cancer diagnosis technique for childhood leukemia disease through the discovery of emerging patterns from the gene expression data, and currently he is very interested in infectious disease studies and water bioinformatics in collaboration with a biological group from the Massachusetts Institute of Technology by exploring graph theories and biological water exclusion principles. One of his data mining articles is widely cited over 450 times, and another paper on bioinformatics is cited over 1000 times, according to google scholar.



Weihang Wang is a master candidate at the computer network information center, Chinese Academy of Sciences. Her research interests include data mining and networking.

Peng Cui is a Ph.D. candidate at the Institute of Zoology, Chinese Academy of Sciences. His research interests include avian behavior and ecology, biodiversity and conservation, as well as avian influenza epidemiology. He is currently working on the bird migration and its potential risk on spread of Highly Pathogenic Avian Influenza H5N1.

Yuanseng Hou is a senior researcher on animal and environment protection in the Bureau of Qinghai Lake National Nature Reserve. He got his bachelor degree from Northwest A&F University, China. He focuses on wild species behavior understanding and has over ten year's field survey experience in the Qinghai Lake, China. This helps him to publish several papers on biological and computation biological journal.

Ze Luo is an associate professor at the Computer Network Information Center, Chinese Academy of Sciences. He got his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2005. His interest includes scientific computing grid, e-Sciences and data mining. He has published over ten papers in international conference and journal.

Jianhui Li is a professor at the Computer Network Information Center, Chinese Academy of Sciences. He got his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2007. His main research interests include large-scale distributed databases management and integration, semantic-based data integration, data intensive computing and scientific application.

Fuming Lei is a professor currently working in the Institute of Zoology, Chinese Academy of Sciences. He is International Ornithological Committee member and vice president of China Ornithological Society. He has published over 140 peer-reviewed papers, including in *Science*, *PLOS ONE, Molecular Ecology, Journal of Virology*, etc. His main research interests include avian systematics, behavior and ecology, biodiversity and conservation, molecular phylogeny and phylogeography, as well as avian influenza epidemiology. He is associate editor-in-chief of *Chinese Birds*, *Acta Zootaxonomic Sinica*, and editorial members of *Current Zoology, Global Journal of Molecular Sciences, World Journal of Zoology, The Open Ornithology Journal,* and etc.

Baoping Yan is a professor and chief engineer in the Computer Network Information Center, Chinese Academy of Sciences. Prior to this, she served as vice president of Dawning Computer Co., a famous computer high-tech company in China. She has completed analysis and design of computer network system, research and implementation of industrial automation and CIMS network technology, ATM-based workstation cluster system, standard management of large-scale network and system integration, Internet/Intranet comprehensive information management system, etc. Currently, she is responsible for the planning and construction of the informatization of the Chinese Academy of Sciences during the 10th Five-year Plan period. She has published over 50 research papers at home and abroad. The government has granted her special allowance for her outstanding contributions.