Exploring the Spatial distribution of Bird Habitat with Cluster Analysis

MingJie Tang1,3, YuanChun Zhou1, Peng Cui2,3, HaiTing Zhang1,2, LiangL Hu1, YuanSheng Hou1, BaoPing Yan1

1Computer Network Information Center, Chinese Academy of Sciences
Beijing, China, 100190
2Institute of Zoology, Chinese Academy of Sciences
3Graduate University of Chinese Academy of Sciences
4Bureau of Qinghai Lake National Nature Reserve
{tangrock@gmail.com}, {yczhou, zht, hll}@sdb.cnic.cn; {cuipeng, j@ioz.ac.cn}, {yb@cnic.cn}

Abstract
Knowledge of wetland use of migratory bird species during the annual life circle is important to construct conservation strategy and explore the implication for avian influenza control. Biological scientists have used GPS satellite telemetry to determine the habitat of wild birds. However, because there is not an efficient method to process the location data sets, scientists have to devote themselves to calculating the scattering location points in the GIS and using MCP or Kernel method to find the approximate home range of bird species.

This paper presents an alternative hierarchy clustering-based methodology to build a Spatial-Tree for discovering the habitat area that overcomes the limit of human observation ability. Preliminary results showed that some nodes of the Spatial-Tree correspond to core range area providing an illusive map to depict the breeding and wintering home range of Bar-headed Goose.

Key Words: Clustering; Bird Migration; Habitat; Scientific data; Qinghai Lake

1 Introduction

The Asian outbreak of highly pathogenic avian influenza H5N1 disease in poultry in 2003 and 2004 was unprecedented in its geographical extent, and its transmission to human being was an ominous sign [1]. Some research findings indicate that domestic ducks in southern China play a central role in the generation and maintenance of this virus and those wild birds may have contributed to the increasingly wide spread virus. This assumption had led to another question: how can we understand the habitat, migration distance and time about this species. Indeed, understanding of this species’ habitat is critical for us to find the roots of some questions, such as how the wild life and domestic poultry intersect together and the possibility of H5N1 spilling over from the poultry sector into wild bird species [2].

Traditionally, the spatial analysis of the specie’s transmission coordinates and the layered maps were performed by GIS (Geographic Information System) including ESRI’S ARCINFO 7.1.2 and ArcView 3.1 (Research Institute, Inc., Redlands, California, U.S.A.). For the lacking of efficient method to process the location plots and the limitation of data, biological scientist have to count those location plots in a certain area and then utilize kernel model [3] to describe the home range of bird species. Therefore, it should be the only way to identify the stop area of the species and the wintering areas. However, until now they have not found a persuasive way to discover much more important information such as the region of the area and the stay duration of birds in habitat. The difficulty lies in the nature of the bird species’ location data is accumulating and the number of plots increased greatly. The situation becomes further complicated when all of this data plots come together in one map. What is more, such sparsely analysis data are excessively dependent on the researcher’s impression.

This paper presents an alternative clustering-based methodology for discovering the habitat which overcomes these limitations. The employ of clustering is driven by the intuition that a potential habitat is expected to involve a significant region of wild bird species that prefer to stay much longer time, and we expect that such a habitat will be the place where the location points will be denser over the entire area. This paper illustrated one hierarchy cluster approach to find the habitat with different density based on DBSCAN [4]. This approach offers a number of benefits: (i) it can discover a habitat in the whole area without the observation. (ii) Core area and out core rang region would be easily viewed in some bigger habitat range (iii) It provides an effective way to manage the discovered habitat from the Spatial-Tree which was built as our hierarchy clustering process.

It’s encouraging to see the results of applying our methodology to discover habitat. Three major habitats were included, namely, breeding place, stop over area and wintering area, which approximately confirming the past observation. In fact, we also can determine the duration of
migration from the cluster area. The preliminary work is to help scientist to determine the risk of infecting wild birds with viruses from poultry and vice versa. Finally, we evaluate the site-fidelity, namely the birds always go back to the same wintering or breeding sites they have visited before. Some habitat shows much more “stronger affective” than others, which is an interesting biological topic and on the other hand reflects our cluster quality.

The organization of our paper is as follows: Section 2 describes the research background and location data. In Section 3, our method of clustering and building Spatial-Tree is discussed. Section 4 presents the results of applying this methodology to find habitat. And in Section 5 we give a summary about our work and figure out our directions in future study.

2 Background and Data

Our studies were conducted at the Qinghai Lake National Nature Reserve, Qinghai province, China. Qinghai Lake is situated in the Qinghai Province, which is the largest salt lake in China, with an area of 525 Km². We complied movement data form 29 bar-headed geese (Anser indicus) from Qinghai Lake. 14 geese of them were captured on Mar25-31, 2007, and the others were captured on March 28 -April 3, 2008. Each bird was weighed, measured and equipped with an 45 g solar-powered portable transmitter terminal (PTT:9 North Star Science and Technology, LLC, Baltimore, Maryland U.S.A; and 1 Microwave Telemetry PTT-100,Columbia, Maryland U.S.A). Transmitter signals were received by Argos data system (CLS America Inc., Maryland, U.S.A) and transmitter locations were estimated. Argos classified the location accuracy into seven categories: 3, 2, 1, 0 and A, 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 this kind of location data LG.

We transform Argos data (from Western Ecological Research Center) from text files into a relation database. The bird migration data for our analysis consists of global snapshots of plots information of variables (e.g., direction, latitude and longitude) for all birds. In this paper, we focus on spatial and temporal information, obtained for each bird every two hours during the satellite traction process.

We have conducted two bird migration surveys on Qinghai Lake in 2007 and 2008. We have collected 266796 and 12951 useful and unique location data for 2007 survey and 2008 survey separately until1 Nov 1, 2008. 90.1% of were of LC 0-3 and LG quality, which would be used in our application for discovering habitat; the remaining LA,B,Z will be dismissed. PTT were deployed on 14 bar-headed geese from Qinghai Lake in March 2007. Three PTTs were still active as of 1 Nov 2008. Our traction has covered the whole routine of bird migration between breeding areas and wintering areas. Only three PTT were lost before the birds return back their wintering place. In addition, among the PTTs that were deployed on 15 bar-headed geese from Qinghai Lake in March 2008, nine of them were still active as of Nov 1, 2008. Until Nov 1 2008 most of them have arrived at winter area.

3 Clustering based methodology for the discovery of home range

In this section, we elaborate on three key steps for clustering the data sets: discovering the cluster by DBSCAN, which is one famous density based cluster algorithm. In addition, we extend the traditional hierarch cluster method based on DBSCAN to rediscover the spatial relationship between clusters as the different parameters in the DBSCAN approach. And the resulting Spatial-Tree was built to describe this core area in the bigger home range of bird. We then come to find the Minimum Convex Polygon Home Range based on the traditional convex algorithm so as to much more precisely reflect the cluster results.

3.1 Related Works and DBSCAN

Cluster analysis is either used as the stand-alone tool to get insight into the distribution of data sets so as to focus on further analyzing or data preprocessing, or as a data preprocessing step for other data analysis works which could be taken later. Various techniques can be used for clustering spatial data sets. For instance, an improved k-medoid method, called CLARANS [5] is introduced. And SNN [6] also was developed to discover arbitrary-shaped clusters.

In this work, we will use DBSCAN as our clustering approach. This method can find clusters of arbitrary shapes. The goal of the algorithm DBSCAN is to partition a database into sets of objects, i.e. clusters, such that the density of objects inside of each cluster is considerably higher than outside of the cluster. Furthermore, the density within the areas of noise is lower than the density in any of the clusters. 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 with a density above a specified threshold, MinPts, are classified as core points, while noise points are defined as non-core points that don’t have a core point within the specified radius.

131
3.2 Hierarchical Cluster and Spatial-Tree building based on DBSCAN

Although we have found suitable parameters of Eps and MinPts, this brings about other problems. If a user defines the neighborhood of a point by specifying a particular radius and looks for core points that have a pre-defined number of points within that radius, then either the tight cluster on the map will be picked up as one cluster, and the rest will be marked as noise, or else every point will belong to a single cluster. We then can assume that the cluster area with small Eps would tend to be the children node for the bigger cluster area.

Also, for MCP range area commonly includes areas never visited by animal. In order to get the actual area used within bar-headed goose home ranges, core areas were calculated. The core area usually defined as areas receiving concentrated by individual at each wetland. For example, the fostering place would be core region and the foraging area would be the out core region. To find the core area in the breeding MCP range area or other bigger ones, we have to run the cluster approach again on this area with much smaller Eps. Meanwhile, there is hundreds of core area scattering on the map. It is necessary to find an efficient and easy way for further management of these results. Motivated by this requirement, we develop a Hierarchical DBSCAN cluster method, which can reflect the relationship between different clusters, where the algorithm can build Spatial-Tree from top (large cluster) to bottom (several clusters), and coding the index for every cluster as the Spatial-Tree growing.

The problem we hope to solve out is how to cluster hierarchically, these methods can be either agglomerative or divisive. Agglomerative methods (eg. AGNES) starts with as many clusters as variables and builds a tree by successive join the two nearest clusters. On the other hand, divisive methods (eg. DIANA ) starts with one big cluster and successive splits the largest cluster by the most dissimilar variable. AGNES and DIANA are fully described in chapters 5 and 6 of [12].

Hierarchical cluster Approach:

Different from the DIANA which select the cluster with largest diameter and divided into two new clusters until all clusters contain only one variable. We utilize the breadth-first search (BFS), in graph theory, BFS is a graph search algorithm that begins at the root node and explores all the neighboring nodes to build the Spatial-tree from top to bottom. Our hierarchical cluster approach employs the similar strategy like the BFS. At first the only one cluster which contains all datasets was selected and DBSCAN with the predefined parameter would split this cluster into sub cluster. Those cluster results were pulled into one queue, and then the data sets from the queue was selected to run DBSCAN again. Our splitting phase will stop until the cluster level reach the defined level from top to bottom. When the cluster was spitted into smaller ones, that cluster was indexed as well. Take the space into consideration we omit a detailed description of the hierarchy algorithm. But we ought to note the following points:

A. We need to initial two pointers: first and last one to identify the tree level in the Queue

B. If the number of cluster result from his parents data node is lower than 2, we will not pull this cluster results into the queue.

C. Members are in the same level of tree, they will share the same cluster parameter Eps and MinPts we predefined before.

D. We use one garbage collector to save the cluster which will be the noise member in the next tree level. Because the parameter will change as the tree level change.

E. As the Spatial-Tree growing, every cluster node will be encoded by their father cluster id joining with its own cluster number

Consider the Fig 3, the depth of Spatial-Tree is four and the cluster area also was encoded. For example: the left most leave node in tree was encoded with 0/0/1/0, thus, we can know its parent node is 0/0/1 and its grandfather is 0/0/0.

Fig 3: Spatial-Tree

The hierarchy cluster procedure has deployed the mainframe to discover the habitat, and then the next step is to define object similarity and cluster parameter for every tree level.

Similarity based on the Geographical distance

As previously stated, only a dissimilarity matrix is needed to perform a divisive cluster analysis, so we should concentrate our efforts on this matrix efficiency. The location points in our paper are geographic coordinates (latitude and longitude), geographical distance formula [7] was deployed to calculate the great-circle distance. In addition, bird migration route overlaps broadly with The East Asian Flyway of Anaitdate and the East Asian-Australasian Flyway of shorebirds[9]. We set the $r$ (sphere of radius) to 6378140 meters in the distance formula $d = \Delta \delta \cdot \cos(b)$. Where $\Delta \delta$ is the central angle, $\delta$ is the geographical distance.

Determining the Parameter of Hierarchy Cluster

We employed the following procedures which could allow us to compute the approximately cluster parameters Eps and MinPts for hierarchy cluster method.

1. In general, we only interest in location points that bird species tend to stay longer rather than the location they fly over occasionally, thus we used the $T_{sum}$, which was associated with the time bird moving in one certain scope of area to define this phenomena. Therefore, the number of MinPts could be computed by formula 2:

$$\text{MinPts} = T_{sum} / T_p \cdot T_p$$

$T_{sum}$ is the time interval which location
signal we received. In our paper, $T_{num}$ is set as two days, because a stopover event as the use of a migration area for more than two days after departure from either the breeding area or the wintering area [9]. And the average time $T_p$ for each bird from our data sets was two hours. Then the resulting MinPts can be 25.

2. To get the parameter Eps, we develop a simple but effective heuristic to determine the parameters Eps. The heuristic is based on the following observation. The wintering and breeding area usually cover at least ten thousand square kilometers wide region. We can assume the Eps would be at most ten thousand meters. Otherwise, if the Eps was much higher, the cluster area would be over estimated and would not promise us as hoped. In addition, we have to reduce the amount of Eps value with multiples of 10 from the 10,000 m, 1000 m, 100 m to 10 m to get a smaller region of space.

3. The stay duration of one wild bird in its own habitat is naturally determined. Look at from another way, the temporal distribution of bird location also can reflect the spatial distribution from our cluster approach. So we utilize the migration time recorded [9] and our observation to rectify the parameter Eps several times to gain our ends.

3.3 Minimum Convex Polygon Home Range

To know the precise range of the habitat region, Kernel density estimation has become prevalent in the home range analysis [3]. In our work, MCP (Minimum Convex Polygon Home Range) would be more proper in our home range estimation [10]. We first go to compute the polygon in this area which would be a convex . There are two algorithms that compute the convex hull of a set of n points. Graham’s scan runs in $O(n \log n)$ time, and the Jarvis’s march runs in $O(n h)$ times, where h is the number of vertices of the convex hull. Both Graham’s scan and Jarvis’s march use rotational sweep, which processing vertices in the order of the polar angles they form with a reference vertex. At first, Point with Maxim or Minimum latitude was computed, because only they will be the point on the convex. And then we utilize Graham-Scan to get the MCP. The run time will be limited to $O(n)$ time. A much more technical description could be seen in [11].

Home range was computed based on spherical geometry from the convex results. A closed geometric figure on the surface of a sphere is formed by the arcs of great circles. The spherical polygon is a generalization of the spherical triangle [12]. If $\Phi$ is the sum of the radian angles of a spherical polygon on a sphere of radius R, then the area is:

$$s = [\Phi - (n - 2)\pi] \times R^2$$

4 Discovery and analysis of cluster area

In this section, we describe the results of discovering habitat on 2007 survey and 2008 survey data sets using our hierarchy cluster approach. Because of space constraints, we mostly focus on the 2007 survey, and some in 2008 survey also were presented. As one part of our discussions, we will analyze site-fidelity, implication for other research works as well.

4.1 System framework of HAMap

We are developing HAMap (Habitat Area Map), a web GIS system based on the Google map which greatly improves a user's Web visualization experience. Fig 4 shows the system architecture of the Cluster Visual Map. Here, Google Map is a popular web based GIS developed by Google. The main tasks of the HAMap are listed as follows: .Our web server will download the latest data from the mail box and then save them to the data base automatically.

When a web Browser input the parameters in our .jsp page (Java serve net page). The host analyzes the input parameter and translates them to the cluster model .cluster model at first run the cluster algorithm on the datasets from database, and then respond the user with the document stream as the .xml type, which are detailed by the different cluster result such as the location information, the area of convex hull and so on.

The xml stream will be sent to Google Server net and the Browser will display the cluster information based on the Google Map.

4.2 Spatial-Tree for bar-head goose

The different annual migration route would not overlap with each other. This involves dividing those data sets into two parts: 2007 survey and 2008 survey. 2007 Spatial-Tree was built from top to bottom on the annual migration data from 2007-03 to 2008-03. In the Fig 5, the left slide is the tree and the right is the spatial distribution associated with that node. The convex home range also was depicted with different color of polygon and the description would be shown when the user clicks the marker with certain index.

When the Spatial-Tree depth is 3, there are 6 clusters in the level 1 with Eps is 35000 meters and MinPts is 25, and the average range of habitats area is 29045.38 Km$^2$. In the map, 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 in China with 9254 Km², and it is interesting to find that one species (No. BH07_67693) moved to cluster with index 1 within Mongolia rather than stay in Qinghai Lake for breeding. The movement of bar-headed goose depicted in this study conformed to the Central Asian Flyway [8]. There are eight species migrated to Tibet river valley, and stay there from 2007-10-21 to 2008-02-25, with a total of 127 days for overwinter rather than fly to north-eastern India and Bangladesh.

The maximum one is the wintering area in Tibet river valley in China with 9254 Km², and it mainly covers the salt lake illustrating the importance of salt lake. These salt lakes are integral to the hydrology defining the landscape and vegetation characteristics of the region.

In the third level of Spatial-Tree, 146 nodes are computed with the Eps 300 meters and MinPts 25. We also focus the cluster result in Qinghai Lake. The core area colored red which indexed as 2 in Fig 7 precisely covers one famous small island named San Kuai Si within the Qinghai Lake, a famous breeding place. Seven bar-head geese have stayed there from 2007-04-10 to 2007-06-24 for 74 days. This illustrated that our cluster method could detect the core range area precisely.

We also can discover the main movements of the fourteen bar-headed geese in 2007 in Fig.5. They depart the breeding place in Qinghai Lake and then arrive at the post breeding area in Zhalin-Eling lake area or Huangheyuan wetland and stay there about two months before heading to south. Then they follow the cluster 2 and cluster 5 which served as the stop over area, and finally arriving at the winter area. Mean fall migration time was about 7 days.

2008 Spatial-Tree was built on the location data from 2008 survey during the period from 2008-03-01 to 2008-11-01 in the Fig 8. The size and distribution of post breeding and breeding area approximately conformed to the cluster result from 2007 survey. This suggests that bar-headed goose did not move randomly during the whole process of migration. At last, seven birds have arrived at the wintering place in Tibet river valley. The fall migration distance is about 1600 km.

4.3 Discussion and Implications for habitat conservation and avian influenza

There are two migration seasons for bar-headed goose: spring migration and fall migration. Biological scientist always aspired to know whether those species would fly back their original habitat in the next year. This is the so called site fidelity. On other hand, the site fidelity will reflect the quality of our cluster area result more or less. Three birds in 2007 survey still active even on Nov 1. We choose the species (NO.BH07_6795) from them and draw its location data from the 2008-04-01 to 2008-11-1 on our 2007 discovered habitat in Fig 9. Those
location points with green color cover most of our discovered MCP in 2007. This not only proves this species would like to return their habitat next year, but also shows the outcomes of our research to some degree.

Fig 9: Site-Fidelity of Spring Migration in 2007

Our cluster based approach for discovering the bar headed goose approximately depicts the geographical distribution of this wild water bird. Both of the cluster results in 2007 and in 2008 matched greatly, which showed us that some certain habitat such as the Qinghai Lake, DaLing Lake and the Tibet river valley are of vital importance for some species to live. The wide area of MCP proves that it is necessary to build one broad network to cover the different core region area. The cluster result displayed in the GIS paves the way for human beings to construct one systematic nature reserve in future. In addition, scientists would do much more research works such as virus, plant and environment quality survey to discover the way of highly pathogenic avian influenza disperse in the wild bird species' MCP.

5 Conclusion and Future Work

Satellite tracking has been used successfully to determine the migration routes and stopover sites of a number of birds. Such information allowed the development of a future plan for protecting the breeding and stopover sites. However, the lack of an efficient data analyzing approach presents an obstacle to help researchers to do this work automatically and precisely. In this paper we suggest to exploit the field of using the location data information as a supplement clustering process which can provide an alternative approach for visual observation in counting the location plots for finding the wild bird species’ habitat. Meanwhile, in order to discover the core rang, a new cluster strategy was proposed: A hierarchy cluster based on the BFS. Our initial analysis of the clustering results suggests that this clustering strategy can effectively manage the different cluster area and discover the spatial relationship of cluster area. It does provide an effective assistance for biological researchers in discovering new habitat.

In the future, we plan to extend our current work to address several unresolved issues. Specifically, we want to figure out if there is any frequent and interesting visiting sequence for different wild birds. One possible solution is to extract the possible candidate sequence based on some famous sequence mining approach. Also, we have begun to investigate comparing the cluster area spatial distribution between different years, and hope to discover the habitat changing trend of bird species. Finally, we intend to extend our analyses to other species in the Qinghai Lake to identify the cross habitat for different species.

Acknowledgements

This work was supported in part by Chinese Academy of Sciences (No.INFO105-5D2) and 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 with us.

Reference