2017 International Conference on Security, Pattern Analysis,
and Cybernetics (SPAC)

Spatial-Temporal Analysis on Bird Habitat Discovery in China

Xiaoming Zhan, Yanming Ye*, Yaoxin Zhuo
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou 310018, China
Corresponding author: yeym@hdu.edu.cn

Benyun Shi*, Yizhi Ren, Weitong Hu
School of Cyberspace
Hangzhou Dianzi University
Hangzhou 310018, China
Corresponding author: benyunshi@outlook.com

***Abstract*—Exploring migration patterns through uncovering migratory birds' habitat information is very important in biology, which has scientific significance in animal habitat conservation and avian influenza control. In this paper, we convert the traditional biology problem into a computational study and use data mining techniques to analyze the spatial and temporal distribution of bird-watching data in China. First, we present an improved hierarchical clustering algorithm (IHDBSCAN) to identify the habitats/stopovers of migrant birds. Then, we use a kernel smoothing method to fit the temporal distribution of bird observation in each spatial cluster. A hierarchical cluster tree is generated where the leaf nodes indicate different bird habitats/stopovers. Finally, the results is visualized on the map of China. Experimental results show that the proposed algorithm can effectively find the spatial and temporal distribution of Anseriformes' habitats.**

I. Introduction

Spatial clustering is a popular data mining method, which groups a set of objects into meaningful subclasses based on certain metrics, such as similarity, distance, and density. In recent years, more and more clustering techniques are proposed and applied in different fields such as machine learning, data mining, image processing, pattern recognition, information retrieval and so on. There are many possible classification for these clustering approaches [1–3]. Most commonly, these clustering approaches can be classified into three basic types: partitioning, hierarchical, density-based methods.

Partitioning algorithm requires an input parameter k, where k represents the number of clusters. For instance, k-means [4], k-medoids [1], and improved k-medoid, called CLARANS [2]. The shape of all clusters discovered by these methods are convex which is restrictive. Furthermore, some prior knowledge is required, which is not available for many applications.

Hierarchical algorithm creates a hierarchical decomposition of the data set and forms a dendrogram. In such hierarchy, each node of the tree represents a cluster. The dendrogram can be created by a bottom-up strategy (Agglomerative algorithm). For example, the AGNES builds a tree by merging adjacent clusters continuously from the initial clusters that only have one object (data) respectively. The dendrogram can also be created by the top-down strategy (divisive algorithm). For example, DIANA builds a tree by separating clusters continuously by dissimilar variables from the initial cluster that contains all the objects (data), specific description in Chapter 3 of [3]. BIRCH [5] uses a CF-tree for partitioning the data sets in an incremental and dynamic way. CURE [6] select a fixed number of well-scattered objects to represent each cluster and then shrinking them toward the cluster centers by a specified fraction. In contrast to partitioning algorithm, hierarchical algorithms do not need k as an input. However, a terminal condition has to be defined which indicating the merge or division process should be terminated.

Density-based algorithm consider clusters as dense regions in the data space which are separated by regions of low density (noise). DBSCAN [7] clusters the data objects based on the density, and is widely used in many fields due to its simplicity, robustness against noise and ability to discover clusters of arbitrary shapes. However, DBSCAN requires two input parameters, Eps (the radius of the cluster) and MinPts (the minimum numbers of neighbours inside the cluster). Then, Ester et al. propose a IncrDBSCAN method [8] based on DBSCAN to process the spatial data sets as well. Later, Ankerst et al. [9] and Daszykowski et al. [10] proposed OPTICS to find the suitable parameters Eps and MinPts in the DBSCAN. Moreover, Karami et al. [11] proposed a BDE-DBSCAN method to quickly and automatically specify appropriate parameters values for Eps and MinPts, the Eps parameter can be calculated by an analytical way as in [12]. In addition, Rodriguez and Laio [13] proposed a clustering method by fast search and find of density peaks.

It is unreasonable to determine the number of habitat beforehand in bird migration, and most of the habitat is a dense areas gathered by individual in general. Thus, the density-based algorithm DBSCAN is a quite good choice. However, perhaps one migration habitat is composed of its subset regions, when people need to discover the subset of large habitat, existing DBSCAN method or other density-method provide no solution to the data sets. Combined with the method in [14, 15], an improved hierarchical DBSCAN algorithm (IHDBSCAN) is presented to adapt to the data sets, which is capable of discovering several significant habitats.

Primary goals of IHDBSCAN algorithm presented in this paper is to discover the most significant habitats [16] in bird migration. The presented algorithm provides a number of benifits: (i) since the precise number of birds migration

978-1-5386-3016-7/17/$31.00 ©2017 IEEE

habitats is unknown, thus the algorithm should be capable to automatically identify the number of clusters rather than manually setting it in advance. (ii) it can filter out those areas with lower density or less counts to discover the significant habitats. (iii) it can reduce the sensitivity to noise, improve the efficiency of processing data. (iv) it is able to select the appropriate radius for clustering according to the density of data, and discover arbitrary-shaped clusters. (iv) it can build the level-tree for providing an effective way to manage the discovered habitat.

We designed the experiments based on the spatial-temporal data to evaluate the efficiency of our algorithm in finding the Anseriformes' habitats, and the results are visualized on the map, meanwhile a kernel smooth method is used to fit the temporal distribution for each cluster. The results are very useful for future study in migration route, virus infection and so on.

The rest of the paper is organized as follows. In section II, we present a statement of the problem. In section III, we describes the characteristics of bird observation data and our algorithm in detail. The experimental results are presented in section IV. Finally, we draw a conclusion and point out our future work in section V.

II. Problem Statement

Discovering habitat information during the bird migration process is an important study in bird migration. The activities of birds in the habitat are more frequent, with more data records that represent denser than other non habitat areas. The paper improves hierarchical DBSCAN algorithm to find a dense area through bird migration data without predefined number of clusters. First, the DBSCAN algorithm is used to cluster the bird migration data with larger radius that will get a wide range of bird stop area, and then in such a wide range of area to DBSCAN clustering is repeated in a top-down manner, and the results are organized in a level-tree to manage clusters. Furthermore, the kernel smoothing method is used to generate the temporal distribution for each cluster obtained by IHDBSCAN.

III. Method

A. Data Collection

Collecting data through citizen science is a popular trend in recent research. The method in [17] is used to collect bird's citizen science data, utilizing a generic framework, data collected by multiple citizen scientists or volunteers can be submitted by mobile phone, together with GPS data, to a common web database. This approach allows two-way communication between the bird observer and the web database, all data can be queried from the database and viewed and analyzed using Google Maps (or Google Earth) both via the web and on the mobile phone. Participants follow a checklist protocol, where time, location, and counts of birds are all reported in a standardized manner.

In this way, samples in database during 2008-2009 in China were extracted as data sets. The data sets is very rich, with about 157,322 observation records, including almost all bird species. These records cover all 34 provinces, municipalities and autonomous regions, including Hong Kong, Macao and Taiwan. Hence, the bird-watching data in China is comprehensive, and it accurately reflects the spatial distribution of birds. In this paper, we adopted Anseriformes as the experimental objects, the detail description about Anseriformes are represented by the form shown in Table I. For example, one location record (Bird ID:64; Latitude:22.8559; Longitude:109.9450; Time:2008-10-03; Birds Num:250) indicate that 250 birds with ID 64 was observed in location (Latitude:22.8559; Longitude:109.9450) at 2008-10-03.

TABLE I
Bird-watching data about Anseriformes in China.

Bird ID	Latitude	Longitude	Time	Birds Num
64	22.8559	109.9450	2008-10-03	250
64	24.7479	97.5599	2008-04-17	17
64	22.8559	108.3404	2008-10-12	1
64	24.6120	97.6567	2009-03-03	3
65	43.7903	87.4497	2008-08-25	19
65	43.7903	87.4497	2008-09-16	17
65	43.7903	87.4497	2008-07-20	7
65	43.7903	87.4497	2008-10-04	5
65	43.7903	87.4497	2009-05-16	3
66	40.7583	107.4267	2009-05-07	26
66	37.3965	118.8091	2008-12-19	270
67	34.8293	111.1941	2008-02-10	670
67	34.7611	111.2334	2008-02-11	460
67	40.4905	116.9558	2008-02-03	400

B. Preprocessing

We are focused on dynamic attributes such as latitude, longitude and time. Then, outlier records are removed. Next, the duplicate data are removed. Note that there are two cases, when irrespective the temporal properties of the data, if two records are observed at the same location, leaving only one of them. When the time information is taken into consideration, if two records are observed at the same location, and the observation time is different, these two records are considered as different observation samples. Finally, all years of data are aggregated into one. Data quality and availability is greatly improved after screening, and subsequent analysis is simplified.

C. Improved Hierarchical DBSCAN (IHDBSCAN)

The aim is to mine clusters from the preprocessed data set. In other words, it is to find significant Anseriformes' habitats with different density based on DBSCAN. Motivated by this requirement, the proposed method is described in detail in this section.

1) Setting DBSCAN parameters: Until recently, the optimization of the two input parameters in DBSCAN has still not formed a uniform approach. The two initial input parameters, namely Eps and MinPts which both have a significant influence on the clustering results. Hence, DBSCAN is very sensitive to the two parameters. The parameter MinPts is eliminated by setting it to 4 for all dataset (for 2-dimensional data) [7].

978-1-5386-3016-7/17/$31.00 ©2017 IEEE

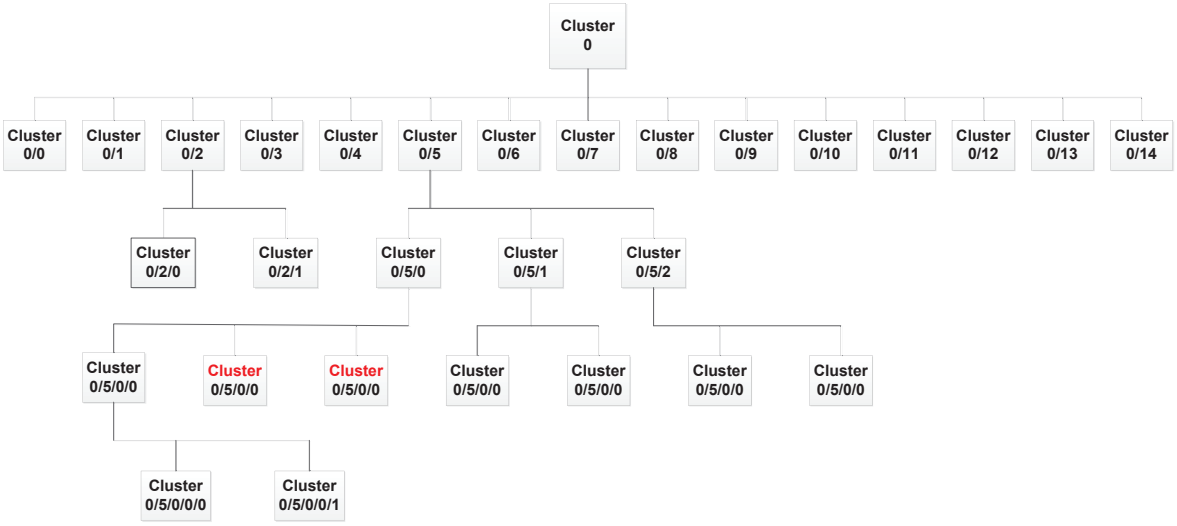

Fig. 1. A level-tree model based on the improved hierarchical DBSCAN method.

The neighbourhood radius Eps is estimated for a data set with the same dimensionality as the data studied, and uniformly distributed within the range of the experimental data. To select an optimal neighbourhood radius for the data set X(m*n), a set of m objects is randomly simulated in the n-dimensional space within the range of the X variables. For example, in this paper, the latitude and longitude of X variables are set to be in China. And then the distances from each simulated object to its kth nearest neighbour are calculated, k being equal to MinPts. The m calculated distances are sorted and Eps is selected as the distance equal to the 95% quantile. As to make the results more accurate, simulation need to be repeated n times. The final Eps is calculated as the mean of n*Eps values.

2) IHDBSCAN algorithm: The two input parameters of DBSCAN are further optimized by a simple and effective approach (see in Section III-C), where Eps need to be re-simulated based on the number of current clusters in each iteration. Then a level-tree is generated which the node indicate the cluster, and the id of a node is jointed by its own cluster label and its father id as the tree grows. By this similar Huffman encoding method, the results are well organized in a tree structure to manage clusters and the cluster id is unique. For example in the Fig.1, the left leaf node in the level-tree is encoded by its father id "0/2" and its own id "/0". Thus, its id is "0/2/0". Note that there are two red marks in the figure, which indicate that the cluster not satisfy the constraint of the algorithm and need to be considered as noise. Furthermore, the depth of level-tree is automatically calculated by IHDBSCAN. In addition, the cluster stops spliting into subclusters should meet the following requirements: (i) the subcluster become all noise or no changes compared to its parent node. (ii) the number of points in each cluster lower than the threshold. The description of the IHDBSCAN algorithm is shown as Algorithm 1.

Algorithm 1 Pseudo code of the IHDBSCAN algorithm

Input: SetOfLocations: X, Distance Matrix: dm, Parameter: MinPts and Eps
Output: X with cluster label

```
queue=[];    //Initialize an empty queue
InitEps=Inf;    //the initial Eps is set to infinity
dm=computeDistance(X);
Root=X;    //root node
queue=Root;    //push object into queue
level=0;   //the height of the tree
while isempty(queue) == 1 do
    for i = 1 → nclust do
        Eps(i)=randomSimulation(cluster(i),dm)
        child(i)=DBSCAN(cluster(i),MinPts,Eps,dm);
  //call DBSCAN
        if childNums==nclust or all child are noise then
            continue;    //cluster unchanged
        else   //subclusters
            queue=[queue;child(i)];
            level=level+1;
        end if
    end for
end while
```

D. Temporal distribution

The time that birds appear in the habitat plays an essential role in bird migration. the kernel smoothing method is utilized to fit the temporal distribution for each cluster, and the key idea of this approach is to smooth the noise as much as possible. The approach is proceed as follows:

(a) The sample data is extracted from the cluster (this data contains time information), and the data is standardized.
(b) A normal kernel smoothing function is used to create one smooth, continuous probability density function for the

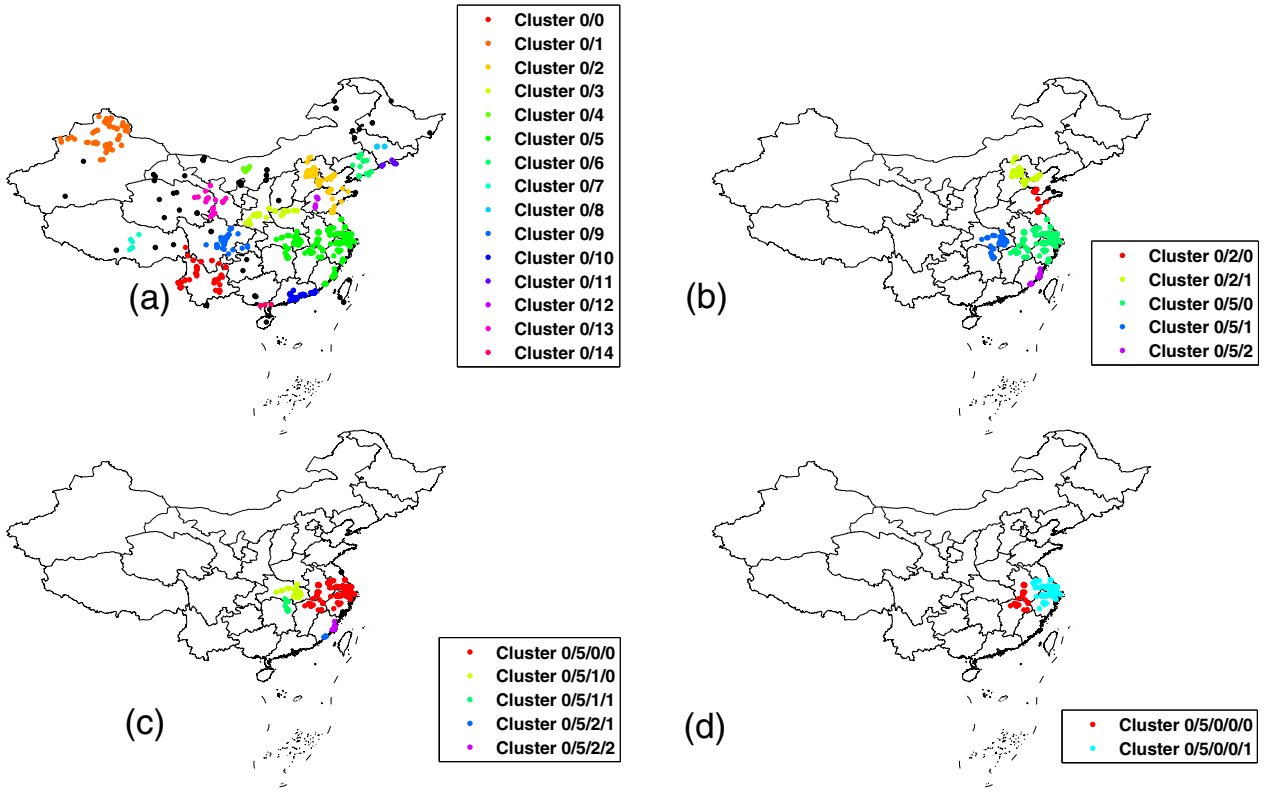

Fig. 2. An overview of the spatial distribution of habitats at different level. Location points with different color represents different migration habitats, while points with black color indicate noise. (a) the first level; (b) the second level; (c) the third level; and (d) the fourth level.

data set, and the parameter support is limited to [0,365].

(c) The mean and standard deviation of the fitted kernel distribution is calculated, and the expression of the fitted distribution is obtained.

(d) Repeat (a)-(c) until all clusters have been calculated.

(e) Draw a smooth curve based on the results.

IV. Results

In this section, the discovered habitats are descripted to evaluate the effectiveness and efficiency of the IHDBSCAN algorithm on the data sets of Anseriformes. This species was chosen because it is more commonly and evenly distributed in China.

1) Discovery of Anseriformes' habitat: The distance between objects is calculated by the spherical distance formula, and the two input parameters Eps and MinPts for DNSCAN were determined as mentioned in Section III. Running the IHDBSCAN algorithm, the depth of the final level-tree is 5, labeled as 0-4. Then, the results are visualized on the map. In the first level of level-tree, there are 15 clusetrs were discovered with the simulated parameters Eps = 152km, MinPts = 4, corresponding to 15 larger habitats. In Fig.2(a), the left panel is the node of level-tree and the right panel is the spatial distribution of habitats associated with these nodes. On the map, location points with different colors represent different clusters, and points with black color indicate noise.

In the second level of level-tree, each node at the first level needs to re-simulated the Eps. As a result, only two larger clusters are split again, and the other clusters remaine unchanged. The nodes with index Cluster 0/2 and Cluster 0/5 at the first level are divide into 2,3 sub nodes with the simulated Eps=142km,139km separately. Cluster index and spatial distribution of habitats is shown in Fig.2(b).

In the third level of level-tree, with the growth of the tree, the simulated Eps is further reduced. Seven subclusters are obtained, but two of them are considered to be noisy because they did not satisfy the restrictions. Removing the two noisy clusters, Fig.2(c) displays the spatial distribution of the remaining five subclusters.

In the fourth level of level-tree, the same procedure is repeated as the previous levels. While one node with index Cluster0/5/0/1 is divided into two subnodes, and the spatial distribution of habitats is shown in Fig.2(d).

2) Temporal distribution of Anseriformes' habitat: The level-tree is generated for clusters management in which the leaf nodes indicate the Anseriformes' habitat discovered by IHDBSCAN in the Fig.1. Traverse the level-tree from top down to get all leaf nodes, and the spatial distribution of all

978-1-5386-3016-7/17/$31.00 ©2017 IEEE

Fig. 3. Spatial distribution of all Anseriformes' habitats discovered by the IHDBSCAN method.

21 Anseriformes' habitats are visualized on the map as shown in Fig.3.

Furthermore, when temporal information for birds in the habitat is taken into cosideration, the frequency and time of birds appearing in each habitat was counted, and a kernel smoothing method is utilized to fit the temporal distribution of Anseriformes' habitat. As hoped, Fig. 4 depicts the temporal distribution of each habitat. For a habitat, it is easy to obtain the largest number of birds for a certain period of time, and the integral of the curve can be calculated according to week or month, i.e. the probability of occurrence, which is helpful for predicting migration routes.

V. Conclusion

The spatial-temporal data of citizen science implicitly record the habitats and migration patterns of birds. However, the location of the habitat, the number of habitats and the relationship between them are uncertain. In this paper, an improved hierarchical DBSCAN clustering algorithm was presented to discover the Anseriformes' habitat. This method can effectively identify the core areas of the habitat with different densities at different levels, and a level-tree is generated to manage different clusters. Also, the kernel smoothing method is utilized to fit the temporal distribution for each habitat which may be useful for future predictive migration routes.

In the future, we plan to address several unresolved issues. First, we need to discover more interesting spatial-temporal patterns [18] and further explore the relationship between habitats. Second, there are major challenges when attempting to apply the two techniques [19], [20] to the problem of bird migration. Motivate by this problem, we intend to use Markov models to modeling the migration problem to infer the migration paths for Anseriformes [21]. In addition, we can extend our analysis to study the main transmission agents and spread route of the virus [22].

Acknowledgment

The authors would like to acknowledge the funding support from National Natural Science Foundation of China (Grant Nos. 81402760, 81573261), and the Natural Science Foundation of Jiangsu Province, China (Grant No. BK20161563) for the research work being presented in this article. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

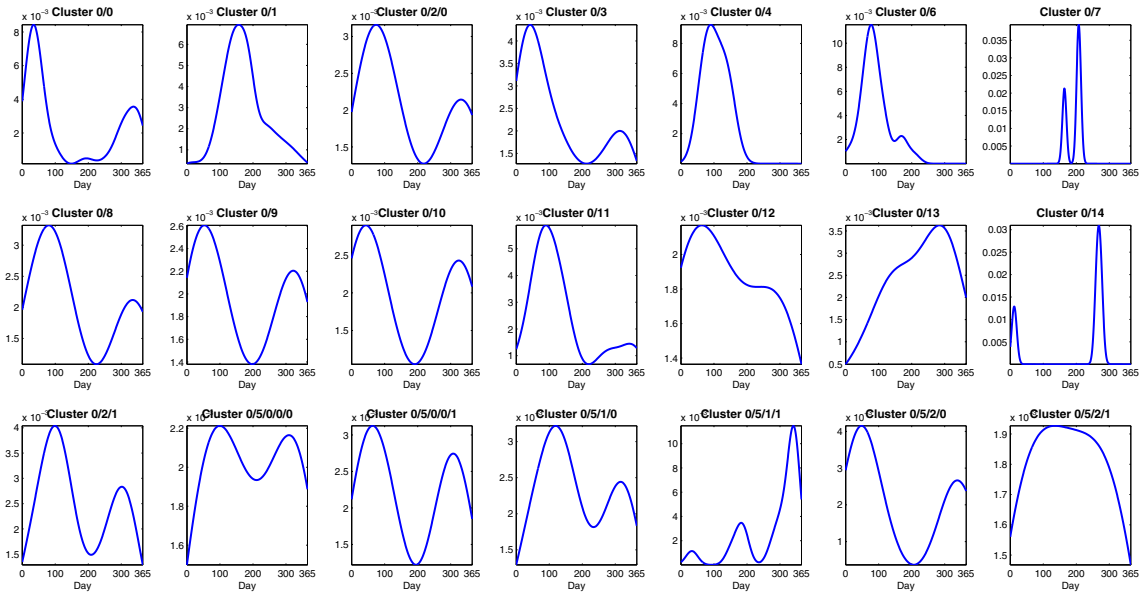

Fig. 4. The temporal distribution of Anseriformes in each cluster fitted by a kernel smoothing method.

REFERENCES

[1] L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, Vol. 344, John Wiley & Sons, 2009.

[2] R. T. Ng, J. Han, Efficient and effective clustering methods for spatial data mining, in: Proceedings of VLDB, Citeseer, 1994, pp. 144–155.

[3] H. Miller, J. Han, Spatial clustering methods in data mining: a survey, Geographic data mining and knowledge discovery, Taylor and Francis.

[4] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA., 1967, pp. 281–297.

[5] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large databases, in: ACM Sigmod Record, Vol. 25, ACM, 1996, pp. 103–114.

[6] S. Guha, R. Rastogi, K. Shim, Cure: an efficient clustering algorithm for large databases, in: ACM Sigmod Record, Vol. 27, ACM, 1998, pp. 73–84.

[7] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, Vol. 96, 1996, pp. 226–231.

[8] M. Ester, H.-P. Kriegel, J. Sander, M. Wimmer, X. Xu, Incremental clustering for mining in a data warehousing environment, in: VLDB, Vol. 98, 1998, pp. 323–333.

[9] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, Optics: ordering points to identify the clustering structure, in: ACM Sigmod record, Vol. 28, ACM, 1999, pp. 49–60.

[10] M. Daszykowski, B. Walczak, D. L. Massart, Looking for natural patterns in analytical data. 2. tracing local density with optics, Journal of chemical information and computer sciences 42 (3) (2002) 500–507.

[11] A. Karami, R. Johansson, Choosing dbscan parameters automatically using differential evolution, International Journal of Computer Applications 91 (7).

[12] M. Daszykowski, B. Walczak, D. Massart, Looking for natural patterns in data: Part 1. density-based approach, Chemometrics and Intelligent Laboratory Systems 56 (2) (2001) 83–92.

[13] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.

[14] M. Tang, Y. Zhou, P. Cui, W. Wang, J. Li, H. Zhang, Y. Hou, B. Yan, Discovery of migration habitats and routes of wild bird species by clustering and association analysis, Advanced Data Mining and Applications (2009) 288–301.

[15] M. Tang, Y. Zhou, J. Li, W. Wang, P. Cui, Y. Hou, Z. Luo, J. Li, F. Lei, B. Yan, Exploring the wild birds migration data for the disease spread study of h5n1: a clustering and association approach, Knowledge and Information Systems 27 (2) (2011) 227–251.

[16] D. B. Neill, A. W. Moore, Rapid detection of significant spatial clusters, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 256–265.

[17] D. M. Aanensen, D. M. Huntley, E. J. Feil, B. G. Spratt, et al., Epicollect: linking smartphones to web applications for epidemiology, ecology and community data collection, PloS one 4 (9) (2009) e6968.

[18] Z. Li, Spatiotemporal pattern mining: algorithms and applications, in: Frequent Pattern Mining, Springer, 2014, pp. 283–306.

[19] D. R. Sheldon, T. G. Dietterich, Collective graphical models, in: Advances in Neural Information Processing Systems, 2011, pp. 1161–1169.

[20] D. Sheldon, T. Sun, A. Kumar, T. Dietterich, Approximate inference in collective graphical models.

[21] M. Elmohamed, D. Kozen, D. R. Sheldon, Collective inference on markov models for modeling bird migration, in: Advances in Neural Information Processing Systems, 2008, pp. 1321–1328.

[22] L. Liang, B. Xu, Y. Chen, Y. Liu, W. Cao, L. Fang, L. Feng, M. F. Goodchild, P. Gong, Combining spatial-temporal and phylogenetic analysis approaches for improved understanding on global h5n1 transmission, PLoS One 5 (10) (2010) e13575.