

# Mining Closed Interesting Subspaces to Discover Conducive Living Environment of Migratory Animals

G.N.V.G. Sirisha and M. Shashi

**Abstract** This paper presents the suitability of subspace clustering techniques to identify the conducive living environment of migratory animals given the geographical and weather conditions prevailing at various locations where the animals thrive. The set of collaborative weather and geographical conditions prevailing at different locations where animals move define the conducive living environment/conditions of animals and hence their accessibility in turn influence the migration behavior of animals. The concept of closed interesting subspaces in density divergence context for multidimensional data is proposed by the authors to model the conducive living conditions of migratory animals. A grid-based subspace mining algorithm namely SCHISM which is originally meant for extracting the maximal interesting subspaces was adapted for finding closed interesting subspaces. Migratory Burchell's Zebra movement data collected from MoveBank was used for this analysis purpose.

**Keywords** Data mining • Subspace clustering • Animal migrations • Closed interesting subspaces • Ecology • Sequence mining

## 1 Introduction

Animal migration is usually a seasonal movement of animals in search of food, suitable breeding sites, or to escape bad weather or other environmental conditions [1]. Migratory animals help us in protecting our food, seed dispersal, pollination,

---

G.N.V.G. Sirisha (✉)

Department of CSE, S.R.K.R. Engineering College, Bhimavaram, India  
e-mail: sirishagadiraju@gmail.com

M. Shashi

Department of CS & SE, Andhra University College of Engineering,  
Andhra University, Visakhapatnam, India  
e-mail: smogalla2000@yahoo.com

© Springer India 2016

S. Das et al. (eds.), *Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015*, Advances in Intelligent Systems and Computing 404, DOI 10.1007/978-81-322-2695-6\_14

and fertilizing the plants [2]. They help in balancing the ecosystem. We need to safeguard them to preserve our ecosystem. Human activities like urbanization, deforestation, overgrazing, agricultural land conversion etc. are leading to habitat loss or habitat fragmentation of migratory animals. This leads to extinction of the migratory animals.

If we are able to identify the favorable environmental conditions and vegetation type for migratory animals, we can reduce habitat loss and habitat fragmentation due to human intervention. It also helps in creating artificial habitats for migratory animals in case of habitat loss to prevent their extinction. Availability of long-term data on migrations annotated with atmospheric observations and landscape variations helps us to analyze the favorable environmental conditions of migratory animals. The ability to locate animals at high spatial and temporal granularities using the GPS and sensor technologies and access to environmental data from different online sources provide rich data which can be analyzed using different statistical techniques, data mining algorithms etc.

Using such data, the authors proposed to apply closed interesting subspace mining techniques to extract patterns that reflect conducive living environment or habitats of migratory animals.

## 1.1 Subspace Clustering

Clustering is the process of grouping objects such that all the objects in each group are similar to one another. In traditional clustering, the similarity between objects is often determined using distance between objects over all the dimensions. Because of the existence of irrelevant and redundant attributes, the objects that are homogenous in only a subset of attributes cannot be detected when seen in full dimensional space [3–5].

Subspace clustering addresses this problem and aims at exploring various subspaces that define clusters specific to such subspaces. The clusters that are existing in subspaces of the multidimensional data space are called subspace clusters. A subspace constitutes the subset of relevant attributes shared by the members of a cluster. The subset of relevant attributes shared by members of one cluster may be different from the subset of relevant attributes shared by members of another cluster.

Let  $\mathbf{D} = \mathbf{O} \times \mathbf{A}$  be a dataset represented in the form matrix, where  $\mathbf{O}$  is the set of objects and  $\mathbf{A}$  is the set of attributes. A subspace cluster  $\mathbf{C}$  is a submatrix  $\mathcal{O} \times \mathcal{A}$ , where the set of objects  $\mathcal{O} \subseteq \mathbf{O}$  is homogeneous in the set of attributes defined by the subspace  $\mathcal{A} \subseteq \mathbf{A}$  [6].

Subspace clustering algorithms discover clusters that exist in multiple, possibly overlapping subspaces, thus allowing an object to be a member of multiple clusters. Subspace clustering techniques can be classified into different types based on the type of data they handle, dimensionality of cluster solutions, and the approaches used for clustering data.

Different subspace clustering algorithms are devised for handling continuous valued data, categorical data, sequence data, stream data, and to provide either 2D or 3D cluster solutions. A 2D cluster solution defines each cluster in two dimensions, with the first dimension representing the objects of the cluster while the second dimension representing the set of attributes shared by the members of a cluster. In other words, each cluster defines a set of objects that are homogenous in a subspace defined by the set of attributes.

It may be noted that the cluster solution given by traditional clustering algorithms is one-dimensional as it groups the objects into clusters in a predefined problem space specified in terms of relevant set of attributes. In 3D data cluster solution, first dimension represents objects, second dimension represents attributes describing the objects, and the third dimension represents an attribute that has to be handled differently like time or location.

Based on the approach taken, subspace clustering algorithms can be classified as grid-based, density-based, and window-based algorithms.

## ***1.2 Animal Migrations***

A lot of research is being carried out to find the factors that influence migration and how migrants respond to changes in climate. In [7], the authors have studied the reasons for adult albatrosses to make long-range trips to preferred, productive areas and how wind assistance facilitates their return flights and how their outbound flights are hampered by head winds.

Altitudinal migration of Galapagos tortoises was studied by Stephen Blake et al. in [8]. The authors studied the roles of environmental variation and individual traits in partial migration of tortoises. A study on migratory behavior of woodland caribou showed that the variability in the movement behavior of woodland caribou is attributed to the local environmental conditions [9] All of the above methods used various statistical techniques to analyze animal movement data.

Li et al. has applied data mining techniques on spatiotemporal data to find periodic movements of animals [10].

Hattie et al. studied the effect of changes in environment on speed and onset of migration of zebras [11]. Hattie et al. have proposed an array of increasingly complex models representing alternative hypotheses regarding the environmental cues and controls for movement. These hypotheses were tested on the movement data of zebras in Botswana. The best models that explained the influence of environmental conditions on departure date and movement speed of migrating zebras were identified. Their analysis has shown that the movement speed of zebras was influenced by Normalized Difference Vegetation Index (NDVI) and rainfall. The date at which zebras started their migration was influenced by cumulative precipitation.

### 1.3 Overview of Proposed Methodology

This research has taken “Migratory Burchell’s zebra in northern Botswana” migratory data [11, 12] annotated with environmental parameters and vegetation indices from MoveBank as a case study to show that closed interesting subspace mining can be used in identifying the conducive living conditions of migratory animals.

Migratory animals move in search of conducive living environment. The climate and vegetation types that are favorable to the development of migratory animals is called conducive living environment. Conducive living conditions detection requires the application of subspace clustering. This is because animal movement is influenced by a subset of atmospheric and landscape variations. We need to identify clusters each of which is a combination of such attributes value pairs (where zebras thrive or move in search of) that defines the conducive living conditions preferred by many zebras. Multiple clusters exist as the atmospheric, vegetative, and geographical conditions defining the conducive living conditions differ with seasons, locations, animal life cycle stage, and their combinations. This requires a clustering algorithm which can identify a cluster that appears significantly and is present in a subset of attributes.

This paper proposes a 2D grid and density-based closed interesting subspace mining algorithm for identifying the conducive living conditions of migratory animals.

SCHSIM [4] a grid- and density-based interesting subspace mining algorithm which mines maximal interesting subspaces was adapted for finding the closed interesting subspaces. A maximal interesting subspace represents the maximal set of collaborative environmental and geographical conditions that appear frequently in the locations where zebras move. A subset of these environmental and geographical conditions may influence the movement of zebras more than the other environmental and geographical conditions defining the maximal interesting subspace. The identification of such strongly influencing factors will be missed if we mine only the maximal interesting subspaces.

So, in multidimensional spaces, the concept of closed interesting subspaces in density divergence context is proposed by the authors. The proposed approach can be used to find conducive living environment of any migratory animal given the atmospheric observations and landscape variations along the path in which migratory animals moved. Though closed interesting subspaces are used to model the conducive living conditions of migratory animals in this paper, the concept of closed interesting subspaces is applicable for identifying interesting subspaces in any multivariate time series.

## 2 Literature Review

There are a number of recent studies for mining interesting subspaces from multidimensional datasets. All of the existing approaches either mine all the interesting subspaces or only the maximal interesting subspaces. Grid-based subspace clustering partitions the data space into grids. Grid cells which have high density are used to form subspace clusters. CLIQUE [13] is a pioneering algorithm in this category. By discretizing the continuous valued attributes, CLIQUE has proposed to apply frequent itemset mining techniques for identifying subspace clusters. It first identifies all interesting subspaces at all dimensionalities. A  $K$ -dimensional subspace corresponds to  $k$ -itemset. The number of objects (support in frequent item mining) that lie in a subspace corresponds to its density. Only those subspaces whose density exceeds a threshold  $\tau$  are mined. Clusters are then discovered by grouping the connected dense subspaces at each dimensionality. ENCLUS [14] uses subspace entropy for selecting interesting subspaces. These interesting subspaces are then used for discovering the clusters. The same clustering model as CLIQUE is used for cluster discovery. MAFIA [15] is a major extension of CLIQUE which uses adaptive grids for cluster discovery.

All the above algorithms use a global density threshold to identify dense units. Because of “density divergence problem,” these algorithms fail to discover high-quality clusters in all subspaces at all dimensionalities. Density divergence problem [4, 16] refers to the phenomenon that cluster densities vary in different subspace cardinalities. As the number of dimensions increase, the data points (database objects) are more spread out in higher dimensional space and they are sparsely distributed. Thus as subspace dimensionality increases, it is constrained by additional attributes or dimensions. Hence, as the dimensionality of subspace increases the number of points in the subspace decreases. If we use the same global density threshold to identify the dense subspaces at all dimensionalities, we may not find high-quality subspaces at all dimensionalities. To overcome this problem, the density threshold value should decrease with increase in subspace cardinality (dimensionality).

SUBCLU [17] uses DBSCAN [18] cluster model of density connected objects to discover the subspace clusters. A cluster is defined as a maximal set of density connected points. It takes two input parameters  $\epsilon$  and  $m$ . SUBCLU uses the same parameter values at all subspace dimensionalities, so SUBCLU may not identify high-quality clusters at all subspace dimensionalities.

FIRES [19] a filter-refinement based subspace clustering algorithm is proposed to overcome the scalability and density divergence problems faced by most density-based subspace clustering algorithms. As it mines only the maximal subspace clusters, some clusters that appear significantly at lower dimensionalities will not be detected.

INSCY [20] and Scalable Density-Based Subspace Clustering [21] are two other efficient density-based subspace clustering algorithms. Both INSCY and Scalable Density-Based Subspace Clustering do not deal with the density divergence problem. DUSC [22] is a density-based subspace clustering algorithm which uses

DBSCAN model for cluster discovery. It uses a density measure that is adaptive to the dimensionality.

DENCOS [16] is a subspace clustering algorithm proposed to solve the density divergence problem. It uses density threshold that adapts to subspace cardinality.

SCHISM [4] is a grid-based interesting subspace mining algorithm that overcomes the density divergence problem. It is a highly scalable algorithm taking an order of minutes to mine subspaces from datasets which are as large as 3 million. Though DENCOS also overcomes the density divergence problem, it uses a non-parametric method in calculating the density thresholds at different subspace cardinalities. SCHISM uses a parametric method to find the density thresholds at different subspace cardinalities using Chernoff-Hoeffding's inequality. It mines maximal subspaces by constraining a subspace with additional dimensions while ensuring the density thresholds are met.

Due to density divergence, the a priori property is violated by those lower dimensional subspaces failing to satisfy their density thresholds even if they are part of a dense subspace of higher dimensionality. Hence, unlike the traditional maximal frequent patterns representing all their subpatterns, a dense maximal subspace may not imply all the subspaces that are subsets of it to be dense. Hence we proposed the concept of closed interesting subspaces in the context of density divergence. Density divergence is inevitable in subspace clustering. This research adapts SCHISM to identify closed interesting subspaces which define the conducive living environment of migratory animals.

### **3 Discovery of Preferred Living Conditions of Migratory Animals Using Density-Based Closed Interesting Subspace Mining**

#### ***3.1 Problem Definition***

Movement trajectories data when annotated with surrounding atmospheric observations and underlying landscape information forms a rich dataset which can be analyzed to find the conducive living conditions of animals.

Many of the previous works in movement ecology have used different statistical models to find effect of changing environment on animal movement. By taking the long-term movement data annotated with atmospheric observations and landscape information, the authors proposed to apply data mining techniques to automatically extract patterns that explain the conducive living conditions of animals.

Migratory Burchell's zebra in northern Botswana data from MoveBank is taken as a case study to show that closed interesting subspace mining can be used in identifying the conducive living conditions of migratory animals. This research aims to find the set of collaborative environmental conditions and vegetation types that attract zebras.

### 3.2 Application of Closed Interesting Subspace Mining on Zebra Data

In multidimensional spaces, the concept of closed interesting subspaces in density divergence context is proposed in this research and SCHISM algorithm was adapted to find the closed interesting subspaces. The overview of the algorithm is given below. It consists of three steps.

Algorithm: Mine Closed Interesting Subspaces from Multivariate Datasets

Input: Multivariate Dataset

Output: Closed Interesting Subspaces

Method:

Step 1: Data Preparation

Step 2: Discretize the database to convert multivariate continuous valued data to discrete valued data

Step 2: Convert the dataset from horizontal to vertical format.

Step 3: Mine closed interesting subspaces.

**Subspace Definition:** A subspace is defined an axis -aligned hyper-rectangle  $[l_1, h_1] \times [l_2, h_2] \times \dots \times [l_d, h_d]$  where  $l_i = (aD_i)/\xi$ , and  $h_i = (bD_i)/\xi$ ,  $a, b$  are positive integers and  $a < b < \xi$ ,  $\xi$  is the number of divisions of an axis,  $d$  is the number of dimensions. If  $h_i - l_i = D_i$ , the subspace is unconstrained in dimension  $i$  whose range is given as  $D_i$ . An  $m$ -subspace is a subspace constrained in  $m$  dimensions, denoted as  $S_m$  [4].

#### Data Preparation.

The dataset used for this study consisted of 7 adult migratory zebras data. The data is recorded from October 2007 to June 2009. Zebra's location is given in terms of latitude and longitude. Location data is annotated with Weekly-Precipitation, Moderate-resolution Imaging Spectroradiometer (MODIS) Land Terra GPP 1 km 8d GPP, MODIS Land Aqua GPP 1 km 8d GPP, MODIS Land Terra GPP 1 km 8d PsnNet, MODIS Land Aqua GPP 1 km 8d PsnNet variables constituting records with 7 attributes. There are 53,793 such records. Except Weekly\_Precipitation, the other variables data is obtained from MoveBank [12]. The original dataset contained many other variables like timestamp, individual zebras identifier, tag type etc. which were removed because they are irrelevant to influence migration behavior.

Weekly-Precipitation gives the weekly cumulative precipitation in the Tropical Rainfall Measuring Mission (TRMM) grid where the location defined by latitude and longitude is present. The rest of the variables indicate the vegetation productivity. For the calculation of Weekly-precipitation timestamp variable is also considered. The trajectory in which zebras moved is covered by 24 TRMM grids.

For all these 24 TRMM grids, daily precipitation is collected from January 1, 2007 to January 1, 2010. Daily precipitation for all these grids in the specified time period is obtained from IRI data library from Columbia University [23]. These daily precipitation values are then used for finding the weekly precipitation. For each TRMM grid, the weekly precipitation is obtained by adding the daily precipitations of the preceding 7 days of the week that ends with the given day. For every location where a zebra has moved, the weekly precipitation is taken as weekly precipitation of the grid at the timestamp in which zebra is located there.

Zebras followed a highly directed movement during migration which can be captured by applying principal component analysis to latitude and longitude. Latitude and longitude attributes are combined to form a new attribute using principal component analysis. This attribute is named as lat-long. Lat-long is the first principal component obtained by applying PCA to latitude and longitude. The first principal component explained 98.3 % variance in the dataset containing latitude and longitude. Prcomp function from R Language is used for finding the principal components. Latitude and longitude are scaled and centered which normalizes them using Z-score normalization before applying PCA. Weights of the variables longitude and latitude in the computation of the first principal component are 0.7071068 and  $-0.7071068$ , respectively.

The range of values of the following 6 variables Lat-Long, Weekly-Prec, MODIS Land Terra GPP 1 km 8d GPP, MODIS Land Aqua GPP 1 km 8d GPP, MODIS Land Terra GPP 1 km 8d PsnNet, MODIS Land Aqua GPP 1 km 8d PsnNet are  $[-2.21778, 1.49096]$ ,  $[0, 110.1495]$ ,  $[0, 0.03395]$ ,  $[3.99e-009, 0.022086]$ ,  $[-0.01248, 0.027102]$ ,  $[-0.00887, 0.014134]$ , respectively. The values of Weekly\_Precipitation dimension follow a skewed distribution. So, the values are transformed by applying  $\log_2$  transformation except 0 value. Most of the values of the attribute are 0 and it is given a separate interval. First interval corresponds to 0 precipitation. The rest of the transformed values are then discretized into 9 intervals using equiwidth binning. The ranges of values of the rest of the variables are discretized into 10 intervals each using equiwidth binning.

### **Data Discretization.**

The intervals that are generated after discretization are analogous to items in frequent itemset mining. Each interval corresponds to an attribute-value pair. The problem of interesting subspace mining is thus analogous to frequent itemset mining. Each interval is a  $1$ - $D$  subspace. A  $k$ - $D$  subspace is a combination of  $k$  intervals from  $k$  dimensions. Each database record (object) corresponds to a transaction and the intervals correspond to items.  $K$ -subspace corresponds to  $k$ -itemset. The number of objects (support in frequent item set mining) that lie in a subspace corresponds to its density. Only those subspaces whose density exceeds the density threshold are mined. These subspaces are called dense subspaces or interesting subspaces



**Data Transformation.**

After discretization the dataset is converted to vertical format, i.e., for each  $I$ - $D$  subspace we store the set of data records (ridset) where the  $I$ - $D$  subspace occurred.

**Mining Closed Interesting Subspaces.**

SCHISM algorithm is adapted to find the closed interesting subspaces. The proposed algorithm uses a depth first search with backtracking. The algorithm first mines  $1$ - $D$  and  $2$ - $D$  subspaces.  $p$ - $D$  subspaces are used in mining the  $(p + 1)$ - $D$  subspaces. This is done by taking a  $p$ - $D$  subspace as input and the set of interesting  $1$ - $D$  subspaces that can be used to constrain the  $p$ - $D$  subspace in an interval of another dimension. Each such  $1$ - $D$  subspace can be used to constrain a  $p$ - $D$  subspace to produce a  $(p + 1)$ - $D$  subspace, only if the number of points (objects) that exists in this new  $(p + 1)$ - $D$  subspace satisfy the density threshold at  $(p + 1)$  dimensionality [4]. This process continues until the algorithm can no longer extend the given subspace. At this stage, the algorithm backtracks to produce new interesting subspaces by extending other dense subspaces. When the algorithm backtracks, SCHISM algorithm is adapted in this research to check if any of the subsets of the dense subspaces are closed. All the closed subspaces are saved.

**Handling Density Divergence.**

In the problem of mining interesting subspaces as the dimensionality of subspace increases it is constrained by additional attributes or dimensions. Hence, a higher dimensional subspace excludes some of the data points covered by lower dimensional subspace. Volume increases with dimensionality and consequently as the dimensionality of the subspace increases the density of subspace decreases which is referred to as density divergence problem. So in order to find all interesting subspaces of different dimensionalities we have to use high-density threshold at low subspace dimensionalities and low-density threshold in high subspace dimensionalities. Therefore, the candidate elimination process based on antimonotonicity property used in frequent itemset mining algorithms can no longer be used in mining the interesting subspaces.

To overcome the density divergence problem, SCHISM sets different density thresholds for different subspace dimensionalities. All the subspaces with same dimensionality will have same density threshold. Chernoff-Hoeffding bound is used in calculating the density threshold at a given dimensionality.

**Closed Interesting Subspace.**

As a natural consequence of density divergence, the density of a higher dimensional subspace is expected to reduce proportionately as a ratio of density thresholds of corresponding subspaces. The concept of closed pattern/subspace is framed by the authors as follows. A dense subspace of  $p$ -dimensions is closed if all of its extensions have their density reduced by more than the expectation in accordance with the ratio of density thresholds in corresponding subspaces.

Let  $S_p$  denote a dense subspace constrained in  $p$  dimensions.

$S_p$  is closed if  $\forall q > p, \frac{\text{density}(S_q)}{\text{density}(S_p)} < \frac{\text{density\_Threshold}(q)}{\text{density\_Threshold}(p)}$  where  $S_q$  an extension of  $S_p$  is a dense subspace constrained in  $q$  dimensions.

Though this research adapted SCHISM to find dense/closed interesting subspaces, the proposed adaptation is equally applicable irrespective of the way density thresholds are fixed.

### 3.3 Density Threshold as a Function of No. of Constrained Dimensions

Density thresholds calculation in the proposed algorithm is done in the same way as SCHISM [4]. In SCHISM density thresholds are fixed in the following way. A  $p$ -subspace is said to be interesting if the ratio of number of points in the subspace to the total number of points in database is greater than the right hand term in (1)

$$\frac{n_p}{n} \geq \frac{E[X_p]}{n} + \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\tau}\right)} \quad (1)$$

With  $E[X_p] = n\left(\frac{1}{\xi}\right)^p$  where  $\xi$  is the number of intervals into which each dimension is discretized.

The interestingness measure  $\text{thresh}_{SCHISM}(p) = \frac{E[X_p]}{n} + \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\tau}\right)}$  is a nonlinear monotonically decreasing function in the number of dimensions  $p$ , in which the subspace  $S_p$  is constrained. To increase the efficiency and effectiveness of finding interesting subspaces the formula for the density threshold calculation is further optimized by Sequeira et al. [4].

## 4 Result Analysis

The dataset consists of 53,793 records described using 6 attributes. The attributes are discretized into intervals as discussed in Sect. 3.2. Each such interval corresponds to an attribute-value pair. A subspace is a combination of one or more such intervals that appear significantly in the dataset. The significance of subspace is given by its density. All the subspaces whose density exceeds the corresponding density thresholds are called interesting subspaces.

Algorithms that handle density divergence set high-density thresholds at low subspace dimensionalities and low-density thresholds at high subspace dimensionalities. Hence, unlike the traditional maximal frequent patterns representing all their subpatterns, a dense maximal subspace may not imply all the subspaces that are subsets of it to be dense. SCHISM being a maximal interesting subspace miner will miss some of the interesting subspaces at low subspace dimensionalities. Hence

we proposed the concept of closed interesting subspaces in the context of density divergence. Density divergence is inevitable in subspace clustering.

To handle the density divergence problem, nonlinear monotonically decreasing threshold is used as the dimensionality of subspace increases. The density threshold is calculated as shown in Eq. 1. Its value depends on  $\xi$  and  $\tau$  for a given dataset. The performance of SCHISM is good for  $\xi$  values ranging from 5 to 15 [4]. So for mining the closed interesting subspaces,  $\xi$  is taken as 10 in the experiments. The performance of the proposed algorithm is tested for  $\tau$  values ranging from  $\frac{10000}{n}$ ,  $\frac{1000}{n}$ ,  $\frac{100}{n}$ ,  $\frac{10}{n}$ ,  $\frac{1}{n}$ ,  $\frac{1}{10n}$ ,  $\frac{1}{100n}$  and so on to  $\frac{1}{1000000n}$  where  $n$  is the number of records. Figure 1 shows  $\log_{10} \tau$  versus number of patterns. Figure 2 shows  $\log_{10} \tau$  versus time in sec.

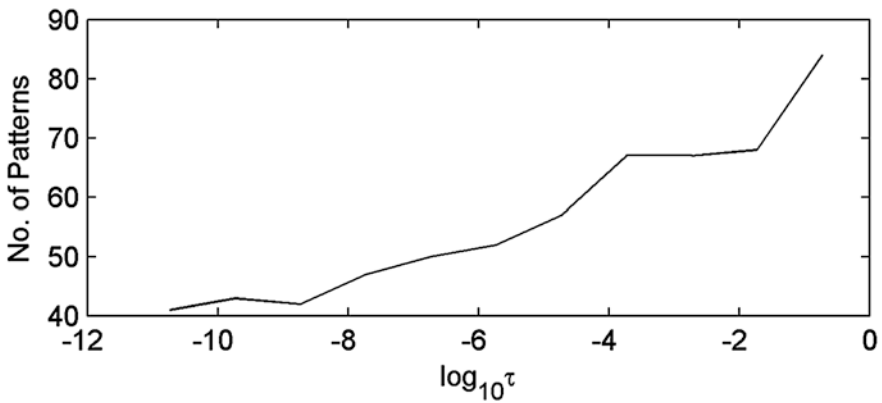


Fig. 1 Number of closed interesting subspaces found for different values of  $\log_{10} \tau$

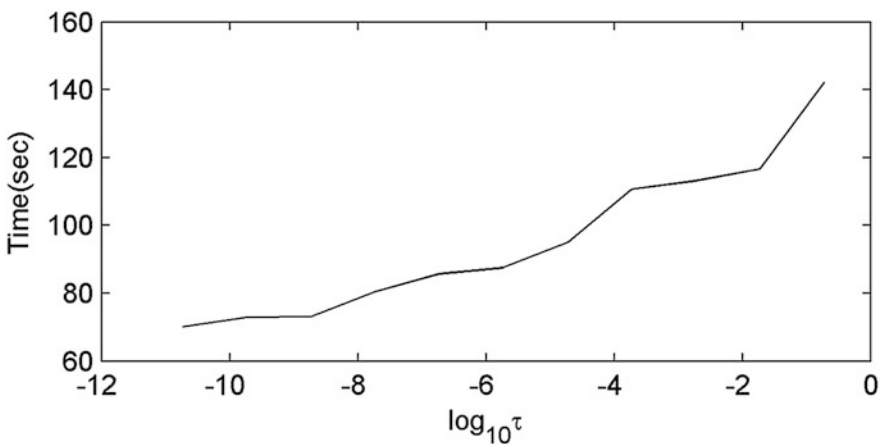


Fig. 2 Time taken by closed interesting subspace miner for different values of  $\log_{10} \tau$

**Table 1** Closed interesting subspaces mined for  $\tau = \frac{1}{n}$ 

Closed interesting subspace	Density	Closed interesting subspace	Density
8 30	2.80892	9 55	5.54347
8 10	6.47482	10 23 45 55	1.51135
17 20 30	2.59699	10 23 55	2.02814
9 17	7.237	10 33 55	2.56725
23 33 45 56	2.71597	10 45 55	3.32757
10 56	3.55622	10 55	6.22386
18 20 30 41 52	2.32186	23 33 45 55	3.19558
18 20 30	3.92058	23 45 55	6.55662
9 18	11.8752	33 45 55	4.48385
0 10 20 30 42 53	2.36834	0 10 20 30 42 52	6.6254
0 20 30 42 53	2.53007	0 20 30 42 52	7.61437
0 10 20 30 53	3.55994	0 10 20 30 42	8.99559
10 20 30 42 53	2.61558	10 20 30 42 52	8.33008
20 30 42 53	4.69206	20 30 42	15.1135
20 30 53	6.92283	9 10 33	1.90545
0 20 30 41 52	2.13225	10 23 33	1.62846
0 20 30 41	3.26065	10 33 45	2.00584
20 30 41 52	6.77783	0 23 33 45	3.26251
20 30 41	10.0013	23 33 45	6.00822
0 23 33 45 55	1.51507	9 10 23	1.61359
0 23 45 55	3.22533	9 23 45	1.70282
0 45 55	4.02097	10 23 45	2.38321
9 10 33 55	1.3775	0 23 45	5.0527
9 33 55	2.50776	9 10 45	1.99468
9 10 45 55	1.31244	9 45	4.90398
9 45 55	2.58212		
9 10 55	3.00411		

The output patterns for  $\tau = \frac{1}{n}$  are shown in Table 1. In the output, literals 0–9 correspond to Lat-long values, 10–19 correspond to Weekly\_Prec values, 20–29 correspond to MODIS Land Terra GPP 1 km 8d GPP, 30–39 correspond to MODIS Land Aqua GPP 1 km 8d GPP, 40–49 correspond to MODIS Land Terra GPP 1 km 8d PsnNet and 50–59 correspond to MODIS Land Aqua GPP 1 km 8d PsnNet, respectively. Interval 60 is used for missing values. Each of the closed interesting subspaces like for example “23 33 45 56” with Density = 2.71597 corresponds to a conducive living environment that appears significantly in the locations where zebras thrive. Each interval in the closed interesting subspace corresponds to an attribute-value pair.

## 5 Conclusions

This research identifies the conducive living conditions prevailing at locations where zebras thrive and hence migrate in search of such conditions. Though many researchers from movement ecology field have done research on effect of environmental variation on animal migration, to our knowledge this is the first work where subspace clustering is applied to the field of movement ecology. The authors integrated the idea of closed patterns with density divergence of subspaces to propose the concept of closed interesting subspaces. SCHISM the grid-based interesting subspace mining algorithm which overcomes the density divergence problem is adapted to mine the closed interesting subspaces. The closed interesting subspaces are used to model the conducive living conditions of migratory animals. The results of this research are used for identifying migratory patterns of zebras as a separate project by the authors, which involve the application of sequential pattern mining algorithms.

**Acknowledgements** Our sincere thanks to Hattie L.A. Bartlam-Brooks for providing access to “Migratory Burchell’s zebra in northern Botswana” data in MoveBank.

## References

1. BBC: Nature migration videos, news and facts. [http://www.bbc.co.uk/nature/adaptations/Animal\\_migration](http://www.bbc.co.uk/nature/adaptations/Animal_migration)
2. Blake, S., Wikelski, M., Cabrera, F., Guezou, A., Silva, M., Sadeghayobi, E., Yackulic, C.B., Jaramillo, P.: Seed dispersal by Galapagos tortoises. *J. Biogeogr.* **39**, 1961–1972 (2012)
3. Beyer K., Goldstein J., Ramakrishnan R., Shaft U.: When is nearest neighbors meaningful?. In: *ICDT*, pp. 217–235 (1999)
4. Sequeira, K., Zaki, M.: SCHISM: a new approach to interesting subspace mining. *J. Bus. Intell. Data Min.* **1**, 137–160 (2005)
5. Kriegel, H.-P., Kroger, P., Zimek, A.: Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering and correlation clustering. *ACM Trans. Knowl. Discov. Data* **3**, 1 (2009)
6. Sim, K., Gopalkrishnan, V., Zimek, A., Cong, G.: A survey on enhanced subspace clustering. *J. Data Min. Knowl. Discov.* **26**, 332–397 (2013)
7. Dodge S., Bohrer G., Weinzierl R., Davidson S.C., Kays R., Douglas D., Cruz S., Han J., Brandes D., Wikelski M.: The environmental-data automated track annotation (Env-DATA) system: linking animal tracks with environmental data. *J. Mov. Ecol.* **1**, 3 (2013)
8. Blake, S., Yackulic, C.B., Cabrera, F., Tapia, W., Gibbs, J.P., Kümmerle, F., Wikelski, M.: Vegetation dynamics drive segregation by body size in Galapagos tortoises migrating across altitudinal gradients. *J. Anim. Ecol.* **82**, 310–321 (2012)
9. Avgar, T., Mosser, A., Brown, G.S., Fryxell, J.M.: Environment and individual drivers of animal movement patterns across a wide geographical gradient. *J. Anim. Ecol.* **82**, 96–106 (2013)
10. Li, Z., Han, J., Ding, B., Kays, R.: Mining periodic behaviors of object movements for animal and biological sustainability studies. *J. Data Min. Knowl. Discov.* **24**, 355–386 (2012)

11. Bartlam-Brooks, H.L.A., Beck, P.S.A., Bohrer, G., Harris, S.: In search of greener pastures—using satellite images to predict the effects of environmental change on zebra migration. *J. Geophys. Res.: Biogeosci.* **188**, 1–11 (2013)
12. Movebank Data Repository: <http://www.datarepository.movebank.org/handle/10255/move.343>, doi:10.5441/001/1.f3550b4f
13. Agrawal R., Gehrke J., Gunopulos D., Raghavan P.: Automatic subspace clustering of high dimensional data for data mining applications. In: ACM SIGMOD International Conference on Management of Data, pp. 94–105 (1998)
14. Chun-Hung. ENCLUS: entropy-based subspace clustering for mining numerical data. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 84–93. ACM, New York (1999)
15. Goil S., Nagesh H., Choudhary A.: MAFIA: efficient and scalable subspace clustering for very large data sets. Technical Report 9906-010, Northwestern University, 1999
16. Chu, Y.-H., Huang, J.-W., Chuang, K.-T., Yang, D.-N., Chen, M.-S.: Density conscious subspace clustering for high-dimensional data. *IEEE Trans. Knowl. Data Eng.* **22**, 16–30 (2010)
17. Kailing K., Kriegel H-P., Kröger P.: Density-connected subspace clustering for high dimensional data. In: 4th SIAM International Conference on Data Mining, pp. 246–256 (2004)
18. Ester M., Kriegel H-P., Sander J., Xu X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231. Portland (1996)
19. Kriegel H-P., Kroger P., Renz M., Wurst S.: A generic framework for efficient subspace clustering of high dimensional data. In: 5th International Conference on Data Mining, pp. 250–257. Houston, TX (2005)
20. Assent I., Krieger R., Müller E., Seidl T.: INSCY: indexing subspace clusters with in process removal of redundancy. In: Eighth IEEE International Conference on Data Mining, pp. 414–425 (2008)
21. Muller, E., Assent, I., Gunnemann, S., Seidl, T.: Scalable density based subspace clustering. In: 20th ACM Conference on Information and Knowledge Management, pp. 1076–1086 (2011)
22. Assent I., Krieger R., Muller E., Seidl T.: DUSC: dimensionality unbiased subspace clustering. In: ICDM, pp. 409–414. Omaha, Nebraska (2007)
23. NASA GES-DAAC TRMM\_L3 TRMM\_3B42 v7 daily Surface Rain from all Satellite and Surface data tables. [http://iridl.ldeo.columbia.edu/SOURCES/NASA/GES-DAAC/TRMM\\_L3/TRMM\\_3B42/v7/daily/precipitation/datatables.html](http://iridl.ldeo.columbia.edu/SOURCES/NASA/GES-DAAC/TRMM_L3/TRMM_3B42/v7/daily/precipitation/datatables.html)