MASTERS PROGRAM IN

GEOSPATIAL TECHNOLOGIES

KNOWLEDGE DISCOVERY FROM TRAJECTORIES

Song Li

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

UNIVERSIDADE NOVA DE LISBOA

UNIVERSITAT JAUME·I

WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

KNOWLEDGE DISCOVERY FROM TRAJECTORIES

Dissertation supervised by
Professor Fernando Bação, Ph.D

Dissertation co-supervised by
Professor Laura Diaz, Ph.D
Professor Miguel Neto, Ph.D

March 2009

ACKNOWLEDGEMENTS

I would like to thank Professor Bação for leading me into this amazing world of data mining. Also many thanks to Professor Painho, the best director I have ever seen, and Professors Michael Gould, Laura Diaz, Miguel Neto for their support, advice and feedback.

Additionally, I will say thanks to Professors Laube, Gennady, Bogorny and Cooley for helping me looking for appropriate data set. Special thank you should goes to Professor Gennady for helping me get access to his software. I have to say his software is amazing. Also thank Robert for helping me to solve all problems related with Geo-SOM. Other thanks are to all the staffs working in ISEGI, your warmth and hospitality make me feel like at home.

Lastly, I want to thank all those passionate researchers—both those cited here as well as others—who are committed to distilling information from trajectory data.

KNOWLEDGE DISCOVERY FROM TRAJECTORIES

ABSTRACT

As a newly proliferating study area, knowledge discovery from trajectories has attracted more and more researchers from different background. However, there is, until now, no theoretical framework for researchers gaining a systematic view of the researches going on. The complexity of spatial and temporal information along with their combination is producing numerous spatio-temporal patterns. In addition, it is very probable that a pattern may have different definition and mining methodology for researchers from different background, such as Geographic Information Science, Data Mining, Database, and Computational Geometry. How to systematically define these patterns, so that the whole community can make better use of previous research? This paper is trying to tackle with this challenge by three steps. First, the input trajectory data is classified; second, taxonomy of spatio-temporal patterns is developed from data mining point of view; lastly, the spatio-temporal patterns appeared on the previous publications are discussed and put into the theoretical framework. In this way, researchers can easily find needed methodology to mining specific pattern in this framework; also the algorithms needing to be developed can be identified for further research. Under the guidance of this framework, an application to a real data set from Starkey Project is performed. Two questions are answers by applying data mining algorithms. First is where the elks would like to stay in the whole range, and the second is whether there are corridors among these regions of interest.

KEYWORDS

Data mining

Elk

Knowledge discovery

Spatio-temporal patterns

Starkey Project

Taxonomy

Theoretical framework

Trajectory

ACRONYMS

BMU – Best Matching Units

CB-SMOT – Clustering-based Stops And Moves of Trajectories

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

FSP – Frequent Sequential Pattern

Geo-SOM – Geo-Self-Organizing Map

GI – Geographic Information

GKD – Geographic Knowledge Discovery

GMT – Greenwich Mean Time

GPS – Global Positioning Systems

GSM – Global System for Mobile communications

KDD – Knowledge Discovery from Databases

LBS – Location Based Services

LCSS – Least Common Sub-Sequence

OPTICS – Ordering Points To Identify the Clustering Structure

RFID – Radio Frequency Identification Tags

RoIs – Regions of Interest

SMOT – Stops and Moves of Trajectories

SOM – Self-Organizing Maps

STARs – Spatio-temporal Association Rules

UTM – Universal Transverse Mercator

ARM – Association Rule Mining

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

1 Introduction

Trajectories are ubiquitous in the real world. As long as an object is moving, no matter it is as huge as a planet or as tiny as an electron, no matter it is a cannonball or a home letter, trajectory is being produced like a long tail. The only difference among them is some movements follow the rule set by Sir Isaac Newton, like guided missile and satellite, while others are under the control of God: God knows where I will go after finishing this thesis. This research is focused on the "unpredictable" trajectories and will start with introducing in the key terms in this context.

1.1 Trajectories

Common sense tells us trajectory is the track of water mark behind a sprinkler vehicle rambling on the street. But a more rigorous definition for scientific research is needed.

In database community, trajectories, which usually stored in moving objects databases (sometimes called trajectory databases in the literature), describe complete histories of movement. A definition giving trajectories semantic meaning was proposed by (Spaccapietra et al. 2008).

"A trajectory is the user defined record of the evolution of the position (perceived as a point) of an object that is moving in space during a given time interval in order to achieve a given goal. Trajectory: $[t_{begin}, t_{end}] \rightarrow$ *space."*

In Geographical Information (GI) science community, people often use the term *Geospatial lifeline*, inspired by time-geography (Hägerstraand 1970; Hägerstrand 1976; Parkes et al. 1980), as a representation of an individual's movement pattern in geographic space.

"A geospatial lifeline is … the continuous set of positions occupied by an object in geographic space over some time period. Geospatial lifeline data consist of discrete space-time observations of a geospatial lifeline, describing an individual's location in geographic space at regular or irregular temporal intervals." (Mark et al. 1998).

Sometimes, tracks or tracking data are used as synonym of trajectories, just like the case of this thesis. Also people often use the term *entity* in place of *object* in above definitions, which often refer to a point object. Typical examples of moving objects under study include vehicles (cars, planes, ships), persons equipped with GPS devices,

animals bearing a transmitter, parcels tagged with RFIDs and hurricane tracking data from meteorological satellites.

The basic element of trajectories is a space-time observation consisting of a triple (ID, Location, Time), where ID, sometimes optional, is a unique identifier of the individual used throughout all recordings of that individual's movements, Location is a spatial descriptor (such as a coordinate pair, a polygon, a street address, or some other locative expression), and Time is the time stamp when the individual was at that particular location (such as a clock time in minutes or event time in years) (Spaccapietra et al. 2008).

In addition, this thesis use the term trajectory instead of spatio-temporal data is to exclude another class of spatio-temporal data set recording spatio-temporal events, such as the records of crime, traffic accident. These events share the same (ID, Location, Time) structure and also applicable to some knowledge discover techniques used in this thesis. But it is beyond the scope of this study.

1.2 Knowledge Discovery

1.2.1 Traditional Knowledge Discovery

Knowledge discovery is often used as a part in the acronym KDD (Knowledge Discovery from Databases), in this case, we are applying it to trajectory databases. KDD is the higher level process of obtaining facts through data mining and distilling this information into knowledge or ideas about the mini-world described by the data. This generally requires a human-level intelligence to guide the process and interpret the results based on pre-existing knowledge (Miller et al. 2001). The KDD process does not seek any arbitrary pattern from a database; rather, data mining seeks only those that are interesting. These patterns are valid (a generalizable pattern, not simply a data anomaly), novel (unexpected), useful (relevant) and understandable (can be interpreted and distilled into knowledge) (Fayyad et al. 1996). The KDD process typically involves the following major steps grouped into larger activity categories (Fayyad et al. 1996; Miller et al. 2001; Qi et al. 2003), which we will also be followed in this study.

1. Background

1.1 Developing an understanding of the application domain; this is often referred to as domain knowledge.

2. Data pre-processing

2.1 Data selection, or determining a subset of the records or variables in the database for focusing the search for interesting patterns.

2.2 Data cleaning, including removal of noise and outliers.

2.3 Data reduction, including transformations, projections and aggregations to find useful representations for the data.

3. Data mining

3.1 Choosing the data mining task. This involves selecting the generic type of pattern sought through data mining; this is the language for expressing facts in the database. Generic pattern types include classes, associations, rules, clusters, outliers and trends.

3.2 Choosing the data mining technique for discovering patterns of the generic type selected in the previous step. Since data mining algorithms are often heuristics (due to scalability requirements), there are typically several techniques available for a given pattern type, with different techniques concentrating on different properties or possible relationships among the data objects.

3.3 Data mining: applying the data mining technique to search for interesting patterns.

4. Knowledge construction

4.1 Interpreting the mined patterns, often through visualization.

4.2 Consolidating the discovered knowledge, either by incorporating the knowledge into a computational system (such as a knowledge-based database) or through documenting and reporting the knowledge to interested parties.

1.2.2 Geographical Knowledge Discovery

Geographic knowledge discovery (GKD) is the process of extracting information and knowledge from massive geo-referenced databases. The nature of geographic entities, relationships and data means that standard KDD techniques are not sufficient (Shekhar et al. 2003).

Geographic data often exhibits the properties of spatial dependency and spatial heterogeneity. Spatial dependency is the tendency of observations that are more proximal in geographic space to exhibit greater degrees of similarity. Proximity can be

defined in highly general terms, including distance, direction and/or topology. Spatial heterogeneity or the non-stationarity of the process with respect to location is often evident since many geographic processes are local.

Including time introduces additional complexity to the GKD process. A simple strategy that treats time as an additional spatial dimension is not sufficient. Time has different semantics than space: time is directional, has unique scaling and granularity properties, and can be cyclical and even branching with parallel local time streams (Roddick et al. 2001).

1.3 Spatio-temporal patterns

As the knowledge discovered from trajectories, the Spatio-temporal patterns should also be valid, novel, useful and understandable. Additionally, due to the innate semantics of spatial and temporal attributes, the number of possible patterns produced by this combination can hardly be exhausted. Roddick and Lees (2001) even argued that the potential complexity of spatio-temporal patterns may require meta-mining techniques that search for higher-level patterns among the large number of patterns generated from spatio-temporal mining. In this section, some common patterns that may be discovered from trajectories are listed as examples, later in this thesis, the patterns documented by researches will be further discussed along with the knowledge discovery techniques, also, a preliminary taxonomy will be proposed.

A class of spatio-temporal patterns is named movement patterns, which in trajectory data, refer to salient events and episodes expressed by a set of entities. In the case of moving animals, movement patterns can be viewed as the spatio-temporal expression of behaviors, as for example in flocking sheep or birds assembling for the seasonal migration. In a transportation context, a movement pattern could be a traffic jam (Gudmundsson et al. 2008) .

Figure 1 illustrated the trajectories of four entities moving over 20 time steps. The following patterns are highlighted: a flock of three entities over five time-steps, a periodic pattern where an entity shows the same spatio-temporal pattern with some periodicity, a meeting place where three entities meet for four time steps, and finally, a frequently visited location which is a region where a single entity spends a lot of time (Gudmundsson et al. 2008) .

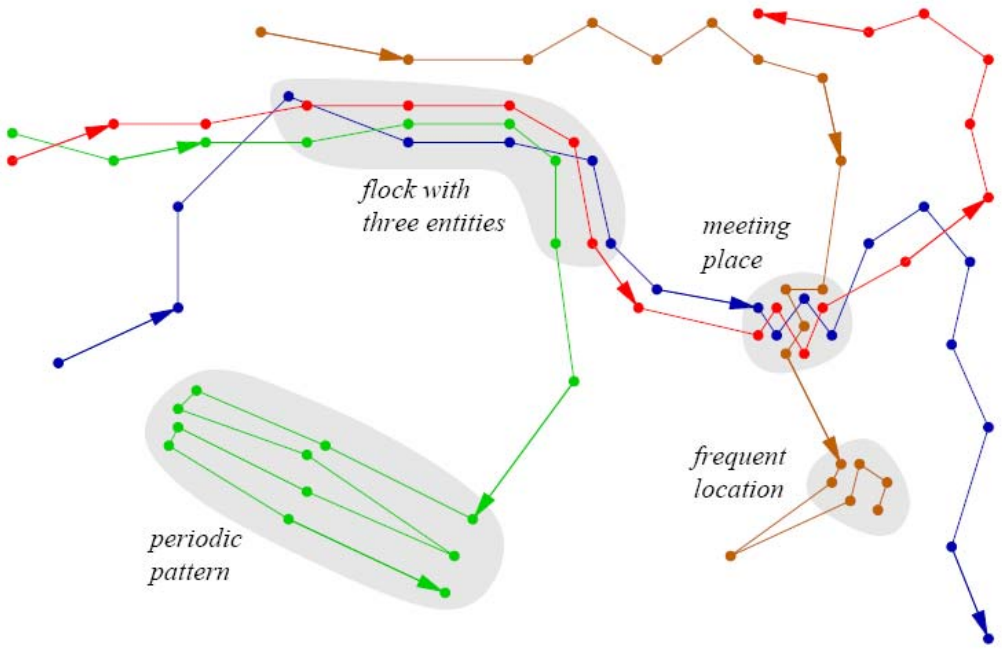

Figure 1: Some common movement patterns (Gudmundsson et al. 2008)

Another class of spatio-temporal patterns is spatio-temporal association rules (STARs). Just like diaper and beer often appear together in the shopping basket, s convenient analogy can be two places in one trip of a moving entity. In practice, this rule is often mined after transforming the trajectory into a series of semantic meaningful places or a series of regions passing by. However, besides using places, other features such as a U-turn in a trip and spatio-temporal attributes of the trajectory can also be used to derive association rules. In addition, a temporal annotation can be attached to the rules, thus the rule can be an object (satisfying some conditions) will show up at place B one hour after going to place A.

1.4 Motivation

The study on trajectory data is attracting more and more attention nowadays, and this booming is motivated by three main driving forces: proliferating data, potential application and challenges facing.

1.4.1 Proliferating data

The availability of trajectory data is the initial driving force and perquisite of related research. Thanks to technical progress on satellite, sensor, RFID (Radio Frequency Identification Tags), video and wireless technologies, especially as mobile devices proliferate and networks become more location aware, large scale capture of the evolving position of individual mobile objects has become technically and economically feasible. The corresponding growth in spatio-temporal data will demand analysis techniques to mine patterns that take into account the semantics of such data.

In addition, for high resolution tracking data, the most distinctive feature is that they allow the track of individuals along an actual movement path, leaving little need for interpretation between sparse observation points. Thus, at almost every instant along the lifeline we can robustly determine the individual's current movement properties, such as speed, acceleration, motion azimuth, path sinuosity, as well as generate even more complex motion properties (Laube et al. 2007).

1.4.2 Potential applications

The availability of trajectory data set opens new perspectives for a large number of applications (from e.g. transportation and logistics to ecology and anthropology) built on the knowledge of movements of objects (Spaccapietra et al. 2008). Some key application fields have been identified by Gudmundsson et al. (2008) as follows:

1.4.2.1 Animal Behavior

The observation of behavioral patterns is crucial to animal behavior science. So far, individual and group patterns are rather directly observed than derived from tracking data. However, there are more and more projects that collect animal movement by equipping them with GPS-GSM collars. For instance, since 2003 the positions of 25 elks in Sweden are obtained every 30 minutes. Other researchers attached small GPS loggers to racing pigeons and tracked their positions every second during a pigeon's journey. It is even possible to track the positions of insects, e.g. butterflies or bees, however most of the times non-GPS based technologies are used that allow for very small and light sensors or transponders. Analyzing movement patterns of animals can help to understand their behavior in many different aspects. Scientists can learn about places that are popular for individual animals, or spots that are frequented by many

animals. It is possible to investigate social interactions, ultimately revealing the social structure within a group of animals. A major focus lies on the investigation of leading and following behavior in socially interacting animals, such as in a flock of sheep or a pack of wolves (Dumont et al. 2005). Advanced path analysis is considered to be a crucial obligation for the interpretation of behavioral experiments conducted with genetically modified animals, for example for water maze experiments with mice exploring spatial learning (Wolfer et al. 2001). Similar analysis is also of increasing interest in agricultural science, contributing to the development, for example, of optimal grazing strategies for cattle with respect to livestock management (Ganskopp 2001). On a larger scale, animal movement data reflects very well the seasonal or permanent migration behavior. In the animation industry, software agents implement movement patterns in order to realistically mimic the behavior of animal groups. Most prominent is the flocking model implemented in NetLogo which mimics the flocking of birds (Wilensky 1998).

1.4.2.2 Human Movement

Movement data of people can be collected and used in several ways. For instance, using mobile phones that communicate with a base station is one way to gather data about the approximate locations of people. Traffic-monitoring devices such as cameras can deliver data on the movement of vehicles. With the technological advancement of mobile and position aware devices, one could expect that tracking data will be increasingly collectable. Although tracking data of people might be available in principle, ethical and privacy aspects need to be taken into consideration before gathering and using this data (Dobson et al. 2003). Nonetheless, if the data is available, it could be used for urban planning, e.g. to plan where to build new roads or where to extend public transport. The detection of movement patterns can furthermore be used to optimize the design of location- based-services (LBS). The services ordered to a moving user could not only be dependent on the actual position, but also on the estimated current activity, which may be derived from a detected movement pattern. Also, the mobile network companies are more effectively allocating resources among wireless cells by predicting next cell in which the mobile user will be.

1.4.2.3 Traffic Management

Movement patterns are used for traffic management in order to detect undesirable or even dangerous constellations of moving entities, such as traffic jams or airplane course conflicts. Traffic management applications may require basic Moving Object Database queries, but also more sophisticated movement patterns involving not just location but also speed, movement direction and other activity parameters.

1.4.2.4 Surveillance and Security Surveillance

Research involving video surveillance (Ng 2001; Porikli 2004; Shim et al. 2003) might have access to more detailed data sets capturing the movement of people, e.g. coordinates from mobile phones or credit card usage, video surveillance camera footage or maybe even GPS data. Apart from analyzing the movement data of a suspect to help prevent further crime, it is an important task to analyze the entire data set to identify suspicious behavior in the first place. This leads to define `normal behavior' and then search the data for any outliers, i.e. entities that do not show normal behavior. Some specific activities and the corresponding movement patterns of the involved moving entities express predefined signatures that can be automatically detected in spatio-temporal or footage data. One example is that fishing boats in the sea around Australia have to report their location in fixed intervals. This is important for the coast guards in case of an emergency, but the data can also be used to identify illegal fishing in certain areas. Another example is that a car thief is expected to move in a very characteristic and hence detectable way across a surveyed car park. Movement patterns have furthermore attracted huge interests in the field of spatial intelligence and disaster management. Batty et al. (2003) investigated local pedestrian movement in the context of disaster evacuation where movement patterns such as congestion or crowding are key safety issues.

1.4.2.5 Military and Battlefield

The digital battlefield is an important application of moving object databases. Whereas real-time location data of friendly troops is easily accessible, the enemy's location may be obtained from reconnaissance planes with only little time lag. Moving object databases not only allow the dynamic updating of location and status of tanks, airplanes and soldiers, but also answering spatio-temporal queries and detecting complex movement patterns. Digital battlefield applications answer spatio-temporal

range queries like "Report all friendly tanks that are currently in region S." A more complex movement pattern in a digital battlefield context would be the identification of the convergence area where the enemy is currently concentrating his troops.

1.4.2.6 Sports Scene Analysis

Research involving sports scene analysis (Moore et al. 2003) are further examples exhibiting deep interests in individual trajectories. Advancements in many different areas in technology are influencing professional sports. For example, some of the major tennis tournaments provide three-dimensional reconstructions of every single point played, tracking the players and the balls. It is furthermore known that, e.g. football coaches routinely analyze match video archives to learn about an opponent's behaviors and strategies. Making use of tracking technology, the movement of the players and the ball can be described by 23 trajectories over the length of the match. Researchers were able to develop a model that is based on the interactions between the players and the ball. This model can be used to quantitatively express the performance of players, and more general, it might lead to an improved overall strategy. Finally, real-time tracking systems are developed that keep track of both players and the ball in order to assist the referee with the detection of the well-defined but nevertheless hard to perceive offside pattern.

1.4.2.7 Movement in Abstract Spaces

In contrast to tracking and analyzing the movement of animals and people on the surface of the earth, it is also possible to obtain and analyze spatio-temporal data in abstract spaces also in higher dimensions. Every scatter plot that constantly updates the changes in the x and y values, produces individual trajectories is open for movement analysis techniques. Two stock exchange series plotted against each other could build such a dynamic scatter-plot. As another example, basic ideological conflicts can be used to construct abstract ideological spaces. Performing factor analysis on referendum data, researchers hypothesized a structure of mentality consisting of dimensions such as "political left vs. political right" or "liberal vs. conservative". Whole districts or even individuals such as members of parliament could now be localized and re-localized in such ideological space depending on their voting behavior and its change over time, respectively. Movement in such a space represents the change of opinions and analyzing this can lead to more insight and understanding of human psychology and politics.

1.4.3 Challenges facing

While there is a growing commitment of resources to the large-scale recording of paths, the analysis commonly conducted with trajectory data remains fairly limited in scope and sophistication (Wolfer et al. 2001). In disciplines outside of geography do not commonly use geospatial methods or theory this may be due to a lack of awareness and understanding of the power of spatial analysis and GI system, and within geography GI science's fetish for the static may be a factor (Raper 2002).

The challenges GI science is facing in developing analytical tools are well documented (Andrienko et al. 2006; Miller et al. 2001) and listed as follows.

a) The spatial relations, both metric (such as distance) and non-metric (such as topology, direction, shape, etc.) and the temporal relations (such as before and after) are information bearing and therefore need to be considered in the data mining methods.

b) Some spatial and temporal relations are implicitly defined, that is, they are not explicitly encoded in a database. These relations must be extracted from the data and there is a trade-off between precomputing them before the actual mining process starts (eager approach) and computing them on-the-fly when they are actually needed (lazy approach). Moreover, despite much formalization of space and time relations available in spatio-temporal reasoning, the extraction of spatial/temporal relations implicitly defined in the data introduces some degree of fuzziness that may have a large impact on the results of the data mining process.

c) Working at the level of stored data, that is, geometric representations (points, lines and regions) for spatial data or time stamps for temporal data, is often undesirable. For instance, urban planning researchers are interested in possible relations between two roads, which either cross each other, or run parallel, or can be confluent, independently of the fact that the two roads are represented by one or more tuples of a relational table. Therefore, complex transformations are required to describe the units of analysis at higher conceptual levels, where human-interpretable properties and relations are expressed.

d) Spatial resolution or temporal granularity can have direct impact on the strength of patterns that can be discovered in the datasets. Interesting patterns are more likely to be discovered at the lowest resolution/granularity level. On the other hand, large support is more likely to exist at higher levels.

e) Many rules of qualitative reasoning on spatial and temporal data (e.g., transitive properties for temporal relations after and before), as well as spatio-temporal ontologies, provide a valuable source of domain independent knowledge that should be taken into account when generating patterns. How to express these rules and how to integrate spatio-temporal reasoning mechanisms in data mining systems are still open problems.

f) Irregularity and Asynchronism: Many types of numerical temporal data are not uniformly and regularly sampled. Also in distributed computing environments like sensor networks, data from different sources might not to be perfectly aligned and hence synchronous methods are inapplicable before rediscretization.

g) Huge volume: The stream of data can be huge for a long, continuous observation period. Many types of measurements can be obtained from a large number of data sources. This requires designing scalable solutions in analyzing a large volume of temporal data, in terms of both the large number of data points and the large number of types of measurements.

h) Exploratory (spatial) data analysis (Anselin 1998) has been identifies as a good way to explore not only spatial but also spatio-temporal data. However, it has also been acknowledged that visual inspection reaches its limits if numbers of moving point objects and lengths of lifelines increase(Kwan 2000).

i) Additional research issues related to spatio-temporal data mining concern data structures used to represent and efficiently index spatio-temporal data.

1.5 Research context and objectives

In face of the challenges brought about by spatio-temporal data. Researchers proposed lots of algorithms to tackle with different data set and to extract all kinds of interesting patterns. However, the theoretical framework for knowledge discovery from trajectories is far from being built. The objective of building such a framework is manifold (Nanni et al. 2008). First, we have to discover the relevant patterns to mine for. Second, taxonomy of these patterns will make it clear for which mining tasks new techniques will have to be developed. Third, suitable algorithmic solutions will have to be proposed to implement these mining tasks. Finally, this new research field could benefit from a clean unified theoretical framework.

Until now the studies in this area are insufficient in many aspects. From the very beginning, the authors often neglect the category of target data sets their methods can be applied to. For example, some methods may be applicable to a single long trajectory, while some methods are only for multiple trajectories. The semantic meaning of discovered patterns also varies correspondingly. Let alone the impact of uncertainty and granularity, which is seldom mentioned in most studies. Secondly, most researches discuss some simple classes of patterns and focuses mainly on algorithmic aspects. These algorithms often neglect the numerous possibilities produced by the special form of trajectories in different semantic context, like the definition of similarity of trajectories. Lastly, the great power of spatial analysis and visualization provided by GI systems is often neglected by researchers from data mining community.

Thus this thesis tentatively classifies all kinds of trajectory data available until now (input) and spatial-temporal patterns (output) respectively. The algorithms (methodology) appeared on papers are put in their applicable range to bridge the data and pattern. Then a real data set is put into this framework to explore the possible patterns and applicable methodology. Standard knowledge discovery procedure is followed along with an emphasis on preprocessing and visualization in GI system. More specifically, the objective of this study will be achieved by fulfilling the four tasks:

- T_1: Classification of trajectory data sets
- T_2: Classification of spatio-temporal patterns corresponding to different trajectory data
- T_3: Bridge the data sets and patterns with methodologies proposed by previous researchers and find out the blanks that should be filled
- T_4: Put a real data set into this framework to explore possible patterns.

1.6 Overview of document

Chapter 1 of this thesis has sought to provide a context to the research objectives by reviewing the relevant terms and listing the motivation. Lastly, this chapter overviewed the objectives of this research, laying out the tasks need to be fulfilled in the following section.

Due to the particularity aspects (e.g. there are very few studies systematically discuss patterns in trajectories) of this research, chapter 2 will not be a traditional literature review, instead, it is a theoretical framework built on related researches. So the details of algorithm are mostly only referred to corresponding papers to avoid diluting the main purpose of this study. As a result, some methods and algorithms used later in the following case study are introduced where they will be applied.

A real data set will be introduced in chapter 3 along with some background knowledge about the data set, and the preprocessing and exploratory analysis will be applied to this data set.

In chapter 4, frequently visited locations, which referred as Regions of Interest (RoIs) are derived from the Starkey data set. Then, after partitioning the trajectories using these RoIs, the trajectories are further clustered to find corridors among these RoIs.

Finally, chapter 5 will offer a concluding summary of this thesis as well as a discussion of suggested areas for future research.

2 Theoretical framework

As illustrated in previous chapter, the theoretical framework is comprised of three parts: trajectories, patterns and methodologies, which will be explored in the following three sections respectively.

2.1 Trajectories

The basic element of trajectories is a space-time observation consisting of a triple (ID, Location, Time). The classification of trajectories then can also be from three aspects: ID, Location, and Time.

2.1.1 ID

ID, the unique identification of location-aware devices bearer, is optional in the trajectory triple if only one moving object is studied. More often case is several or tens of objects are studied, which can produce more complex also more interesting patterns. Additionally, the typical bearers are mainly from three categories: human, animal and vehicle. Others like movement of hurricane landfall (Lee et al. 2007) rarely appears are not classified here. As an incomplete conclusion, table 1 shows a matrix of combination of these two factors and lists a few studies, not exhaustively, on each class of trajectories. No study on single trajectory of animal can be found, and there maybe two reasons explaining this. First, biologists are generally more interested in social behavior and spatial distribution, secondly, patterns shown in many animals as a group, like the Alzheimer-like pathology in laboratory mice (Kritzler et al. 2007), are more persuasive.

	Human	**Animal**	**Vehicle**
Single	(Alvares et al. 2007; Ashbrook et al. 2003; Palma et al. 2008)		(Andrienko et al. 2007)
Multiple	(Ashbrook et al. 2003; Laube et al. 2006; Morzy 2007)	(Carneiro et al. 2008; Kritzler et al. 2007; Laube et al. 2005; Laube et al. 2006; Shoshany et al. 2007)	(Andrienko et al. 2007; Lee et al. 2008; Li et al. 2007)

Table 1: Classification of trajectory data based on moving objects

2.1.2 Location

At first, objects can be moving in free space (birds, ships in the ocean), confined space (cows in pasture, football players in the field) or networks (trucks in road network, ships in river network). The border of free space and confined space can be obscure some

time, since in some sense, everything is moving in a confined space. The difference between then can be determined by the influence of border on movements. The football field is often too small for the 22 players while the ocean and sky is generally big enough for ships and birds. For the same reason, the movement of animal in pasture can be regarded as in free space if the range is big enough. Additionally, in some cases, due to the limitation of tracking technology, movement in free space can be tracked at only a few key places, which comprises a network instead (Kritzler et al. 2007).

Secondly, the locations can be of several formats. Most typical format of location is coordinates logged by GPS devices, which is a pair of latitude and longitude in WGS84 reference system. Another extensively studied format is that of data collected by mobile network using GSM (Global System for Mobile communications), which can only locate the mobile phone holder to a cell ranging several kilometers. Thus the data recorded is only sequences of different cells. The coordinate's pairs can be rediscretized into sequences of cells, actually researchers did in many cases, by dividing the space into grids or only extract important places along the trajectories. A tradeoff between accuracy and mining efficiency is the reason behind the transformation.

A special case is the movement in abstract information space (Laube et al. 2006), where the coordinates can be arbitrary attributes instead of the latitude and longitude we often use, as a result, the border is proprietary to the attributes used.

An inevitable issue related to the location is uncertainty, which is an inherent characteristic of spatiotemporal data (Nanni et al. 2008). It arises due to physical and technical limitations during data collection and storage. Uncertainty of location varies with the applied technology between a few meters (GPS) and kilometers (GSM). In addition, the sampling rate possesses a great influence on accuracy. The faster an object move, to sustain a given level of spatial uncertainty, the more often must an object's location be reported. Background knowledge as well as certain assumptions about movement behavior helps to reduce the uncertainty in data. For example, when tracking a vehicle, we can safely assume that all movements are restricted to the street network. Cars are unlikely to move through buildings.

A frequently made assumption is that of linear movement between two reported positions. The confidence on this assumption is severely influenced by the sampling interval, i.e. the temporal granularity. In general, given two consecutive positions P_1 and

P_2 at times t_1 and t_2 and a maximum speed, an object's position at each moment in time $t \in [t_1,t_2]$ is restricted to some area (Pfoser et al. 1999). If no further information is given, a uniform distribution of the objects within this area can be assumed.

Table 2, similar to table 1, shows the classification along with some studies. Some studies appear both as coordinate pairs and sequence of cells because there are rediscretization as mentioned above.

	Free space	**Confined space**	**Network**
Coordinate pairs	(Alvares et al. 2007; Ashbrook et al. 2003; Palma et al. 2008)	(Moore et al. 2003)	(Andrienko et al. 2007)
Sequence of cells	(Ashbrook et al. 2003; Cheng et al. 2003; Giannotti et al. 2007; Song et al. 2006)	(Moore et al. 2003)	(Alvares et al. 2007; Kritzler et al. 2007)

Table 2: Classification of trajectory data based on format of locations

2.1.3 Time

Time is another special dimension besides location, which can be regarded as ratio, ordinal or cyclic variable in subsequent knowledge discovery process. Here in the classification of trajectory data, only temporal granularity need to be considered. The sampling interval can vary from seconds(Andrienko et al. 2007), hours(Rowland et al. 1997) to days (Carneiro et al. 2008). Also many types of numerical temporal data are not uniformly and regularly sampled. In distributed computing environments like sensor networks, data from different sources might not to be perfectly aligned and hence synchronous methods are inapplicable before rediscretization.

For high resolution tracking data, the most distinctive feature is that they allow the track of individuals along an actual movement path, leaving little need for interpretation between sparse observation points. Thus, at almost every instant along the lifeline we can robustly determine the individual's current movement properties, such as speed, acceleration, motion azimuth, path sinuosity, as well as generate even more complex motion properties(Laube et al. 2007).

The classification of high and low temporal granularity is quite difficult and often depending on the accuracy requirements and objects tracked. The dilemma for trajectories is that often high resolution is required for extracting all important sites, while low resolution is enough for locations between these sites.

2.2 Spatio-temporal Patterns

The spatio-temporal patterns appeared on previous literatures are quite a lot. In this study, we are trying to apply a new taxonomy on these patterns. Our classification extends traditional methods, which are according to data mining techniques, by further classification considering output target and semantic meaning of space and time.

After applying the traditional classification method, i.e. considering the data mining techniques. The patterns are then classified as to the target these patterns are applied to. The target can be ID (a terrorist whose trajectory is unmoral or a bunch of people sharing some interest), or location (a frequently visited area or where traffic congestions often happen), or time (the time when football players slow down or birds start migration). It can also be some spatio-temporal phenomena, for example, the gridlock characterized by large amount of cars moving at low speed.

In the last step, these patterns are further classified according to their semantic meaning of space and time, i.e. the implicit spatial and temporal relations considered. For example, in periodical patterns, the time is regarded as cyclic, while other patterns often treat time as linear.

2.2.1 Classes, Clusters and Outliers

Clusters and classes are basically the same in the sense of similarity between trajectories. In practice, clustering is more often used since predefine the classes is often difficult. In addition, after defining these classes or extracting all the clusters, the leftover trajectories can be regarded as outlier.

2.2.1.1 ID

At first look, this class of patterns only applies to multiple trajectories produced by multiple moving objects; while actually, a single trajectory can be divided into many subtrajectories and given a sub-ID. For example, a person's trajectory can be divided into 1) from home to kindergarten, 2) from kindergarten to office, 3) from office to shopping mall, 4) from shopping mall to office, 5) from office to kindergarten and 6) from kindergarten to home and so on.

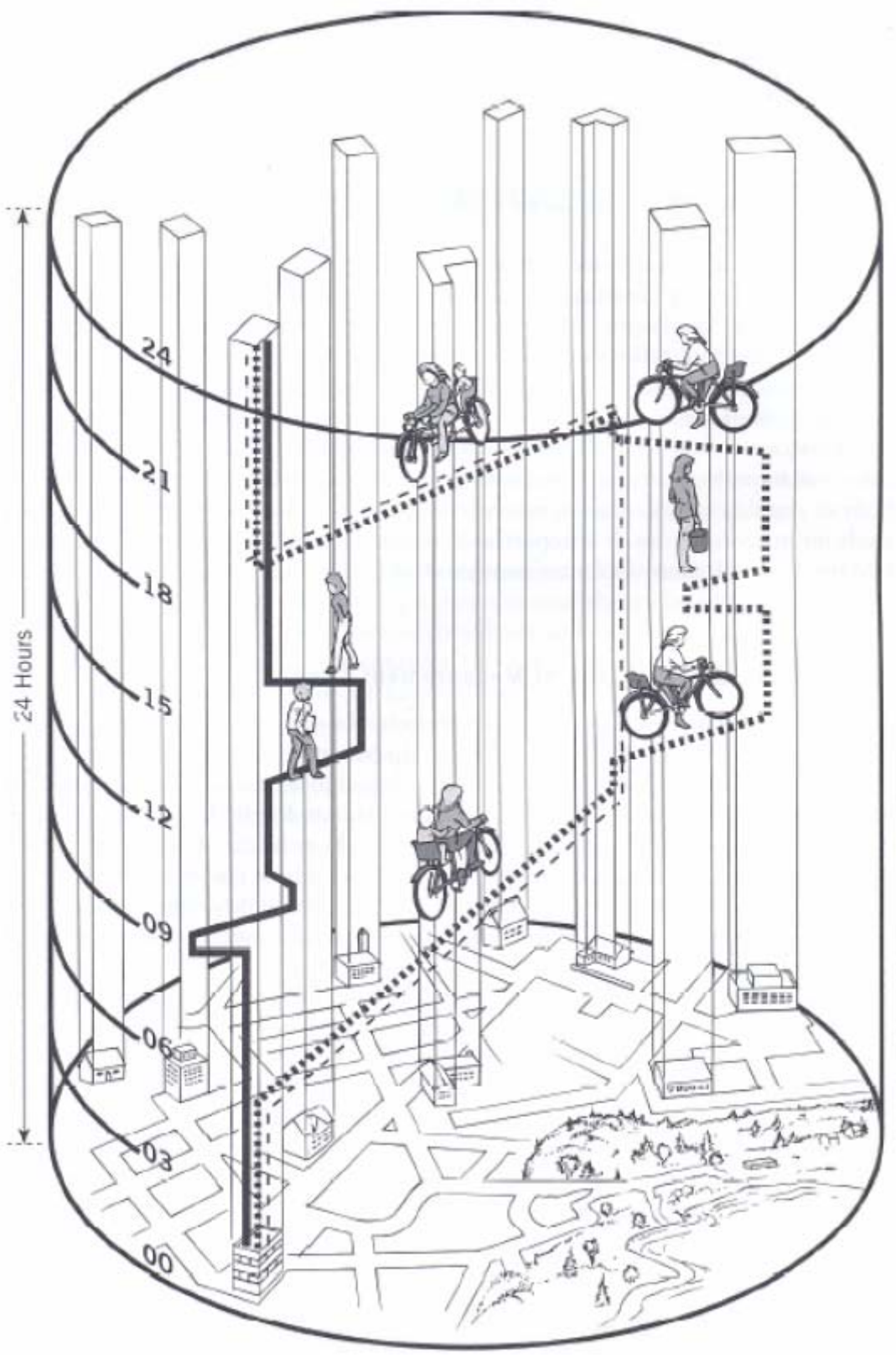

Figure 2: Space-time diagram showing residents in their daily lives (Chrisman 2002)

Just as the example in figure 2, the women's daily movement can be classified into the six categories listed above. As we can see, the definition of these classes is far from exhausting the complexity of spatio-temporal classes. Let's take the class "trajectories from home to kindergarten" as an example. This class can be extended from both spatial and temporal dimension.

At spatial dimension, there are two controlling points in the original definition, while the number of controlling points can be from 1 to positive indefinite. In the case of 1, a class can be defined as "trajectories leaving home". In the case of more than 2, a class can be "trajectories from home to kindergarten passing by a shopping mall. By continuing adding control points, the definition is keeping narrowing down until trajectories whose shape is exactly the same are considered. In practice, it's impossible to use indefinite number of controlling points, thus how to select controlling points, and how the points are distributed along the trajectories are all interesting topic to be discussed.

An implicit assumption above is that similarity of trajectories are based on the distance between controlling points, no matter the distance is metric (like Euclidean) or non-metric (like travelling time). A special but very useful variation is using semantics meaning of these control points instead of physical meaning. In this way, all trajectories in figure 3 are from their home to office, thus all will be classified together. However, until now, there is no exploration in this field due to the difficulty in semantics.

Since coordinates are innate interval attribute, the spatial translation and rotation can be applied if we only want to find cluster of trajectories of similar shape. For example, the trajectories of race cars can be clustered, and then players' performance at different shape curves can be compared. This extension often applies when massive controlling points are used. Also, to avoid the influence of noise, i.e. some controlling points may highly deviate from the right track due to data collecting errors, the extraordinary distance between controlling points can be discarded using voting strategy or probabilistic modeling technique(Gaffney et al. 1999). Certainly, spatial distortion is mathematically feasible; nevertheless, no real life application is conceived.

The temporal dimension, which is completely ignored in the original class definition, can be extended by treat time as an attribute of different measurement scales. When the time is treated as a ratio variable, the controlling points become spatio-

temporal. As a result, in figure 3, the trajectories will be classified into three classes if more than 2 controlling points are considered, even though all trajectories from group 1, 2 and 3 are following the same route; i.e. only trajectories sharing the same route at the same time are regarded as in the same class. When time is treated at interval scale, i.e. the temporal translation is allowed, trajectories following the same route, but the time arriving at each controlling point differs at a given value. For example, group3 and group2, which are parallel in the spatio-temporal space, in figure 3 will be put into the same class. When time is treated as ordinal, i.e. only the sequence of locations is considered. The trajectories in group 1, 2 and 3 will all be recognized in the same class since the same route is followed. A special feature of time is it can be cyclic, like the yearly migration of birds and daily movement of human and vehicles. In this case, the time is ratio (can also be interval or ordinal) but temporal translation at given intervals is allowed. These intervals are often (1day, 2 days, 3 days …) or (1year, 2 years, 3 years …).

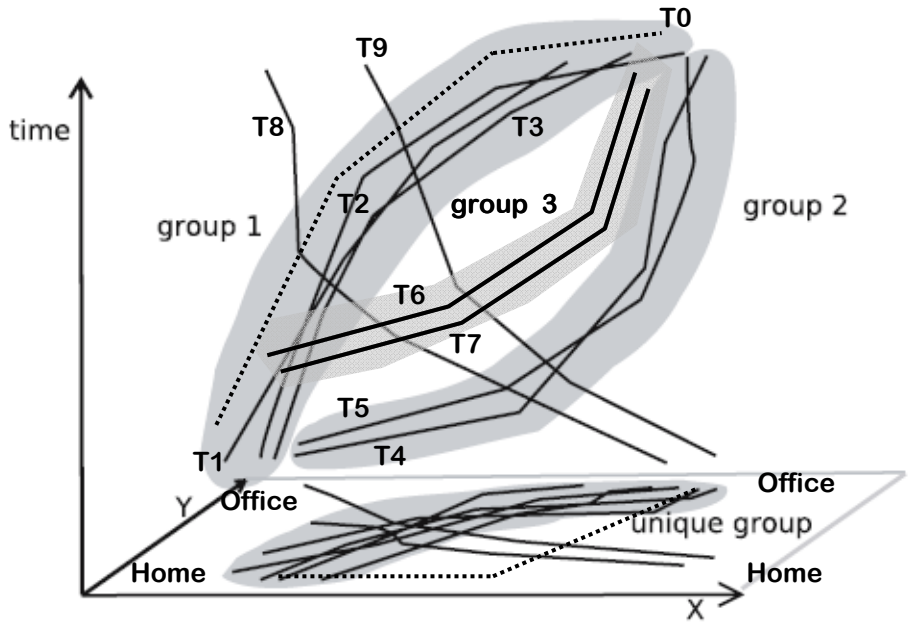

Figure 3: Groups of trajectories, edited from (Nanni et al. 2008)

A special extension of clustering is at ID dimension. This extension leads to a special pattern named moving clusters. As a nominal variable, ID is downgraded into a constant in the patterns in previous analysis. In figure 4, the notable cluster still exists at the three snapshots, but the members of this cluster keep changing. At last, only one

object from the first snapshot stays in this cluster after two periods. Identifying these interesting clusters and studying their temporal evolving is of interest to many biologists.

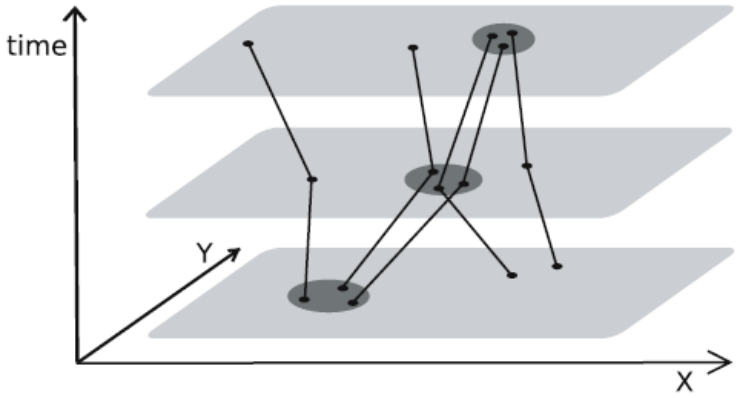

Figure 4: Example of moving clusters (Nanni et al. 2008)

As a conclusion, the classification results from different class (cluster) patterns are listed in table 3 using the example trajectories in figure 3. The ordinal time is not applicable when only one controlling point is considered. The periodical patterns, corresponding to cyclic time cannot be found in this figure due to the difficulty in visualization. Which kind of spatial and temporal extension should be used to define classes and clusters at completely up to the knowledge we want to discover, for example, if we want to find all the (sub-)trajectories of somebody going home from office after finishing work, a two controlling points ratio plus cyclic extension is proper, because he may go home from office in the morning or at noon to fetch something forgotten bringing to office, which is not what we want.

	Ratio	**Interval**	**Ordinal**	**Cyclic**
One physical controlling points(take the starting position of T1)	(T0, T1, T2, T3, T4, T5); (T6, T7); (T8); (T9)	(T0, T1, T2, T3, T4, T5, T6, T7) ; (T8); (T9)	N/A	
Two physical controlling points (starting and finishing points)	(T0, T1, T2, T3, T4, T5); (T6, T7) ; (T8); (T9)	(T0, T1, T2, T3, T4, T5, T6, T7) ; (T8); (T9)	(T0, T1, T2, T3, T4, T5, T6, T7) ; (T8); (T9)	
Several to massive physical controlling points	(T0); (T1, T2, T3); (T4, T5); (T6, T7) ; (T8); (T9)	(T0); (T1, T2, T3); (T4, T5, T6, T7); (T8); (T9)	(T0); (T1, T2, T3, T4, T5, T6, T7); (T8); (T9)	periodical patterns
Semantic controlling points (in this case, only two points, home and office, have semantic meaning)	(T0, T1, T2, T3, T4, T5, T6, T7, T8, T9)	(T0, T1, T2, T3, T4, T5, T6, T7, T8, T9)	(T0, T1, T2, T3, T4, T5, T6, T7, T8, T9)	

Table 3: Class and Cluster patterns

2.2.1.2 Location and time

The classes and clusters for location can be applied in two cases. The first one is using the result from previous section, i.e. a class or cluster of trajectories. Then the space and location occupied by these trajectories will be derived. For example, if trajectories between two semantically meaningful locations are identified as a cluster or class, their spatial projection will show the routes often used, which can be a corridor between animal habitats.

This also the case how classes and clusters are projected to time dimension and the time interval of interest can be derived.

The other case is to find cluster or class of locations visited by the trajectories, in this way, Regions of Interest (RoIs), i.e. frequently visited places, will be identified. The original definition of RoIs, i.e. considering only the frequency of being visited, can be extended at ID and temporal dimension.

When the ID of locations in a cluster is restraint, for example, RoIs can be defined as places which have been visited by 10% of total population of studied moving objects, the RoIs produced will be places of public interest, such as city council, shopping mall and so on. On the other hand, if RoIs is defined as places which have been frequently visited by one moving object, home of each person will be included.

The temporal extension is to cluster or classify at certain snapshot or time span. This extension is necessary because the RoIs in many cases are dynamic. For example, night pub is of interest at night. For animals, they will often spend their day and night at different locations.

2.2.2 Spatio-temporal Association Rules

Association rules are first defined for basket data, which typically consists of the transaction date and the items bought in the transaction. The most famous example of association rule is "on Thursdays and Saturdays males who buy diapers also buy beers".

2.2.2.1 Association rules

To be adapted for mining association rules, researchers often rediscretize trajectories into sequences of cells crossed by trajectories. In this way, a cell is an analogy to an item in shopping basket. The cells can be innate from positioning technique (GSM) or produced by putting uniform grids on study area. Obviously, the transformation to

uniform grid is a uniform information loss from the original trajectory, a better solution is keep the information at important locations and discard unimportant information at other places. Thus researchers usually look for the important locations (home, office, kindergarten…) in the trajectories using clustering technique mentioned above. Then the trajectories are open to all kinds of association rules mining techniques, such as Apriori (Agrawal et al. 1994), FP-tree(Han et al. 2004). An easy example to illustrate this association rule can be "people appear at Lumiar Residence often show up in the Universidade Nova de Lisboa". The support is the proportion of people that appears both at Lumiar Residence and Universidade Nova de Lisboa in the total population investigated, and the confidence is the proportion of people appears both at Lumiar Residence and Universidade Nova de Lisboa in the total population of people appear at Lumiar Residence. This pattern obviously involves both the moving objects (people) and locations (Lumiar and University). Similar to the classes mentioned above, this association rule can be extended both spatially and temporally.

Firstly, we must notice the information loss in the rediscretization, not only in accuracy aspect, but also the implicit topological relations. As a result, the spatial rule "football players whose distance to the ball is less than 10 meters will move towards the ball for at least 2 meters" can never be found after this rediscretization. Also in this example, we can notice that not only the places can be as an analogy to items in shopping basket, a segment of trajectory can also be. The trajectory segments "approaching the ball for at least 2 meters" are physically different, but semantically the same item. This kind of association rules until now has not been explored at all.

Given the rediscretization, the direct application of association rules mining algorithms also ignored the topological relations among these cells. A special case is the application of Markov Chains or Hidden Markov Model, which is often not regarded as association rules. In this model, the possibility of objects moving from one cell to its neighboring cells is actually a spatial confidence, also the topological relation, touch, is considered.

Similar temporal extension for the above mentioned classes and clusters can be applied to association rules. At first, if the time is treated as ordinal variable, the association rules become a specific class named Frequent Sequential Pattern (FSP) mining. The example mentioned above will be modified to "People appear at Lumiar

Residence later on often show up in the Universidade Nova de Lisboa", i.e. a temporal adverb "later on" will be added on. There have been numerous studies to detect FSP in time series data(Agrawal et al. 1995; Pei et al. 2001). Secondly, if the time is regarded as interval attribute, the sequential pattern will be given a fix time interval. In the example above, "later on" will be replaced by "in a hour". In this way, only people show up in one hour in the university after leaving Lumiar will be considered. Thirdly, if the time is a ratio attribute, i.e. no temporal translation is allowed, the example rule above will be "people appear at Lumiar Residence at Feb. 1st, 2009, 9:00 AM often show up in the Universidade Nova de Lisboa at Feb. 1st, 2009, 10:00 AM". In addition, if the cyclic feature of time is considered with a period of 1 day, the rule will be "people appear at Lumiar Residence at 9:00 AM often show up in the Universidade Nova de Lisboa at 10:00 AM every day."

Lastly, we also notice that to characterize a trajectory, discretized regions is not the only and best method, candidates can be used as replacement include:

- Spatial events (visiting some pre-defined spatial regions or visiting twice the same place).
- Spatial, temporal and spatio-temporal attributes (A simple example involving basic aggregation values can be "Length (trajectory) > 50km ⇒ average speed (trajectory) > 60km". Other non-spatial attributes sometimes can also be taken into account.)
- Spatiotemporal events (temporally localized maneuvers like performing U-turns, abrupt stops, sudden accelerations or longer-term behaviors like covering some road segment at some moment and then covering it again later in the opposite direction) as in sequences of the form.
- A segment of trajectory, like the example "approaching the ball for at least 2 meters" mentioned above.

The opposite alternative to the approach above to mining frequent patterns consists in directly analyzing trajectories, for instance to discover paths frequently followed by cars in the city centre, frequent maneuvers performed by animal predators or hunted preys, etc. That means, in particular, that no a priori discretization or other form of pre-processing of spatial and/or temporal information is performed, and therefore the spatio-temporal semantics of data can potentially play a role in the mining

phase. A first consequence of this scenario is that the standard notion of frequent pattern borrowed from transactional data mining, i.e. a pattern that exactly occurs several times in the data, usually cannot be applied. Indeed, the continuity of space and time usually makes it almost impossible to see a configuration occurring more than once perfectly in the same way, and thus some kind of tolerance to small perturbations is needed (Nanni et al. 2008).

2.2.2.2 Location and Time

Apart from the specification of association rules, pattern mining depends on whether the specific focus of the task is on finding interesting patterns or on finding occurrences of the patterns (i.e. where and when they occur and who they involve, which is given the term "Occurrence retrieval"). To some extent, the first case corresponds to direct searches, while the second case corresponds to inverse searches, though the distinction is not crisp. In a direct search, we may specify the hypothesis space H, the space of all patterns regarded in our search, which is usually very large, and aim at identifying all frequent or in another sense interesting patterns $h \in H$. Alternatively, we could specify a set of interesting patterns (or hypotheses) H in advance, H usually being relatively small, and ask for all occurrences that match such patterns in our data (Nanni et al. 2008).

Take the example "people appear at Lumiar Residence often show up in the Universidade Nova de Lisboa" we used before, in the previous section, our intended result may be the rule itself, but in the case of occurrence retrieval, our purpose is to find this kind of location pairs, i.e. "Lumiar Residence" and "Universidade Nova de Lisboa". If this rule is extended to temporal dimension, we can also find the interesting time interval we want. In practice, a user may already have some specific pattern in mind and ask for all of its occurrences, which is basically a synoptic query. The adjective synoptic used here is to differ from elementary query tacked by database community.

2.3 Mining methodology

In this section, we will explore the algorithms developed for the two major classes of spatio-temporal patterns: 1) Classes, clusters, and outliers 2) Spatio-temporal association rules. Then some patterns, such as leading, convergence, which seems difficult to be classified will be given a different interpretation from data mining point of view.

2.3.1 Classes, clusters, and outliers

Since the classification is seldom studied and outlier detection is a byproduct from classification or clustering, we will focus on clustering of trajectories, which is also the most extensively studied area.

2.3.1.1 Traditional clustering method

For trajectory data, to some extent, the definition of the distance or dissimilarity is more important than choosing a clustering algorithm. Considering the classification used in previous section, we will start from the strictest distance measurement, i.e. massive controlling points, treating coordinates and time as ratio attributes, and then gradually extend this distance to more diverse dissimilarity definition.

A simple way to model this strictest distance is to represent trajectories as fixed-length vectors of coordinates and then to compare such vectors by means of some standard distance measure used in the time-series literature, such as the Euclidean distance (the most common one) or any other in the family of p-norms. An alternative solution is given in (Nanni 2002), where the spatial distance between two objects is virtually computed for each time instant, and then the results are aggregated to obtain the overall distance, e.g. by computing the average value, the minimum or the maximum.

Loosening the temporal restriction, two methods from time-series literatures can be applied. One is the comparison of pairs of time series by allowing (dynamic) time warping (Berndt et al. 1994; Vlachos et al. 2003), i.e. a non-linear transformation of time, so that the order of appearance of the locations in the series is kept, but possibly compressing/expanding the movement times. Another method, proposed in (Agrawal et al. 1995) and further studied in (Bozkaya et al. 1997), consists in computing the distance as length of the least common sub-sequence (LCSS) of the two series, essentially formulated as an edit-distance problem.

Another step in loosening the constraints imposed to clusters consists in not requiring a strict co-location of trajectories/routes, but only asking to group objects that perform similar movements, like going in the same direction or performing the same turns (i.e. turns of the same amplitude, whatever the absolute direction). The first example can be simply modelled by defining as similar any couple of objects that follow approximatively the same path but allowing spatial translation, as proposed in (Vlachos et al. 2002) through a translation-invariant, non-metric extension of the above-

mentioned LCSS. A step further is then accomplished in (Vlachos et al. 2004), where a distance that is also rotation-invariant is proposed.

The last loosening is to reduce the number of controlling points. For instance, we could extract all pairs of consecutive values in each series (in our context, consecutive locations within each trajectory) and then simply count the number of pairs shared by the two series compared, as proposed in (Agrawal et al. 1995); or, as an alternative, we could extract a set of landmarks for each time series (i.e. local behaviors of the time series such as minima or maxima or, more specific to our context, changes of speed or direction) and compute the distance between the series by simply comparing their corresponding series of landmarks, as described in (Perng et al. 2000). Another commonly used distance is only considering the starting and ending points to cluster trajectories which are likely having same purpose. A more complex case is to allow a limited amount of random noise in trajectories. Gaffney and Smyth (1999) proposed a mixture model based clustering method for continuous trajectories, which groups together objects that are likely to be generated from a common core trajectory by adding Gaussian noise. In a successive work (Chudova et al. 2003), spatial and (discrete) temporal shifting of trajectories within clusters is also considered and integrated as parameters of the mixture model. Another possible solution can be modify the methods proposed in (Nanni 2002) by introducing a voting algorithm instead of using the sum, average, minimum or maximum value of distances between corresponding points.

Some of the most interesting spatio-temporal patterns are periodic patterns, e.g. yearly migration patterns or daily commuting patterns, which can be regarded as a special clustering when time is treated as cyclic attribute. Mamoulis et al. (2004) considered the special case when the period is given in advance. They partition space into a set of regions which allows them to define a pattern P as a τ-length sequence of the form $r_0, r_1 \ldots, r_{\tau-1}$, where r_i is a spatial region or the special character *, indicating the whole spatial universe. If the entity follows the pattern enough times, the pattern is said to be frequent. However, this definition imposes no control over the density of the regions, i.e. if the regions are too large then the pattern may always be frequent. Therefore an additional constraint is added, namely that the points of each subtrajectory should form a cluster inside the spatial region.

An important issue influencing the results of clustering is the partitioning of trajectories beforehand. Since it is often very difficult to divide the trajectories to proper scale we want, clustering algorithms considering sub-trajectories are also developed (Hwang et al. 2005; Lee et al. 2007; Li et al. 2004; Nanni et al. 2006). The main idea behind these algorithms is to divide trajectories into piece-wise linear, possibly with missing segments. Then, a close time interval for a group of trajectories is defined as the maximal interval such that all individuals are pair-wise close to each other (w.r.t. a given threshold). Groups of trajectories are associated with a weight expressing the proportion of the time in which trajectories are close, and then the mining problem is to find all trajectory groups with a weight beyond a given threshold. If the trajectories are divided according to time intervals, the final result can be the time interval result in the clusters of best quality of clustering and these clusters.

2.3.1.2 Cluster detection using speed

The method mentioned above are all using the coordinates, an alternative is by considering speed, an innate spatio-temporal variable. Laube et al. (2004) defined a collection of spatio-temporal patterns based on direction of movement and location, e.g. flock, leadership, convergence and encounter, and they proposed algorithms to compute them efficiently. These patterns are defined based on their previous study on REMO (Laube et al. 2002). The basic idea of the analysis concept is to compare the motion attributes of point objects over space and time, and thus to relate one object's motion to the motion of all others. The REMO concept (RElative MOtion) is based on two key features: First, a transformation of the lifeline data to a REMO matrix featuring motion attributes (i.e. speed, change of speed or motion azimuth); second, matching of formalized patterns on the matrix (Fig. 5).

The REMO concept allows construction of a wide variety of motion patterns. See the following three basic examples:

- Constancy: Sequence of equal motion attributes for r consecutive time steps (e.g. deer O1 with motion azimuth 45° from t2 to t5).
- Concurrence: Incident of n MPOs showing the same motion attributes value at time t (e.g. deer O1, O2, O3, and O4 with motion azimuth 45° at t4)
- Trend-setter: One trend-setting MPO anticipates the motion of n others. Thus, a trend-setter consists of a constancy linked to a concurrence (e.g. deer O1

anticipates at t2 the motion azimuth 45° that is reproduced by all other MPOs at time t4)

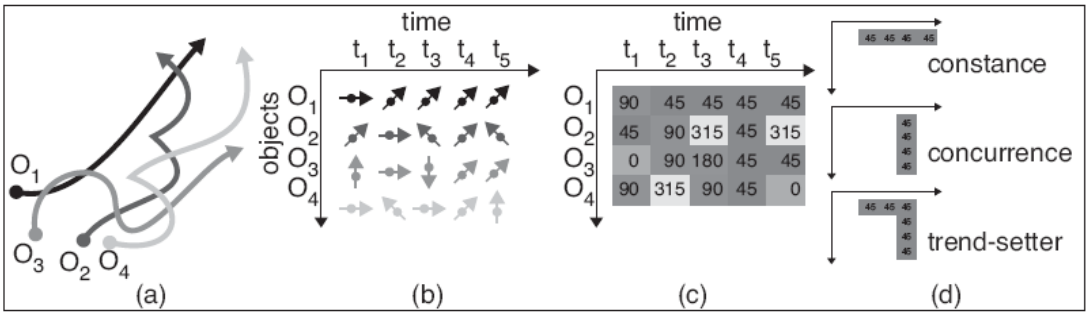

Figure 5: The geospatial lifelines of four MPOs (a) are used to derive in regular intervals the motion azimuth (b). In the REMO analysis matrix (c) generic motion patterns are matched (d).

However, many sheep moving in a similar way is not enough to define a flocking pattern. We expect additionally that all the sheep of a flock graze on the same hillside. Formalized as a generic motion pattern we expect for a flocking the MPOs to be in spatial proximity. Adding spatial constraints to the list of basic motion patterns in figure 2, amended REMO patterns are listed as follows (Fig. 6).

- Track: Consists of the REMO pattern constancy and the attachment of spatial constraint. Definition: constancy + spatial constraint S.
- Flock: Consists of the REMO pattern concurrence and the attachment of a spatial constraint. Definition: concurrence + spatial constraint S.
- Leadership: Consists of the REMO pattern trend-setter and the attachment of a spatial constraint. For example the followers must lie within the range (∂x,∂y) when they join the motion of the trend-setter. Definition: trend-setter + spatial constraint S.

In addition, Laube et al. (2004) proposed the movement pattern convergence (Fig. 7) to answer questions similar to "Can we identify points of interest attracting people only at certain times, events of interest rather than points of interest, losing their attractiveness after a while?" Also noted that entities move towards the same location does not mean they will actually meet there. Thus another movement pattern encounter was also defined.

Convergence: Heading for R. Set of m MPOs at interval i with motion azimuth vectors intersecting within a range R of radius r.

Encounter: Extrapolated meeting within R. Set of m MPOs at interval I with motion azimuth vectors intersecting within a range R of radius r and actually meeting within R extrapolating the current motion.

The opposites of the above described patterns are termed divergence and breakup. The latter term integrates a spatial divergence pattern with the temporal constraint of a precedent meeting in a range R.

Benkert et al. (2007) modified the original definition of a flock to be a set of entities moving close together during a time interval. Note that in this definition the entities involved in the flock must be the same during the whole time interval, in contrast to the moving cluster definition by Kalnis et al. (2005). Benkert et al. (2007) observed that a flock of m entities moving together during k time steps corresponds to a cluster of size m in 2k dimensional space. Thus the problem can be restated as clustering in high dimensional space. To handle high dimensional space one can use well-known dimensionality reduction techniques. There are several decision versions of the problem that have been shown to be NP-hard, for example deciding if there exists a flock of a certain size, or of a certain duration. The special case when the flock is stationary is often called a meeting pattern.

Andersson et al. (2007) gave a more generic definition of the pattern leadership and discussed how such leadership patterns can be computed from a group of moving entities. The proposed definition is based on behavioral patterns discussed in the behavioral ecology literature. The idea is to define a leader as an entity that (1) does not follow anyone else, (2) is followed by a set of entities and (3) this behavior should continue for duration of time. Given these rules all leadership patterns can be efficiently computed.

Among the above mentioned patterns, leadership, convergence and divergence seem difficult to be merged into our classification at first look. However, they are all clustering problems. Leader of a flock can be detected by using clustering algorithm twice. First time, a clustering allowing temporal translation is applied to the trajectories, thus clusters of moving objects following similar routes are generated. Second time, clustering on time dimension is applied to each resulting clusters from first step. Then the leader can be regarded as an outlier temporally ahead of each flock, which is the cluster produced in second step. Convergence and divergence can be regarded as cluster

of trajectories considering only one controlling point, the ending point and starting point respectively. The difficulty now is the extrapolation needed for detecting convergence.

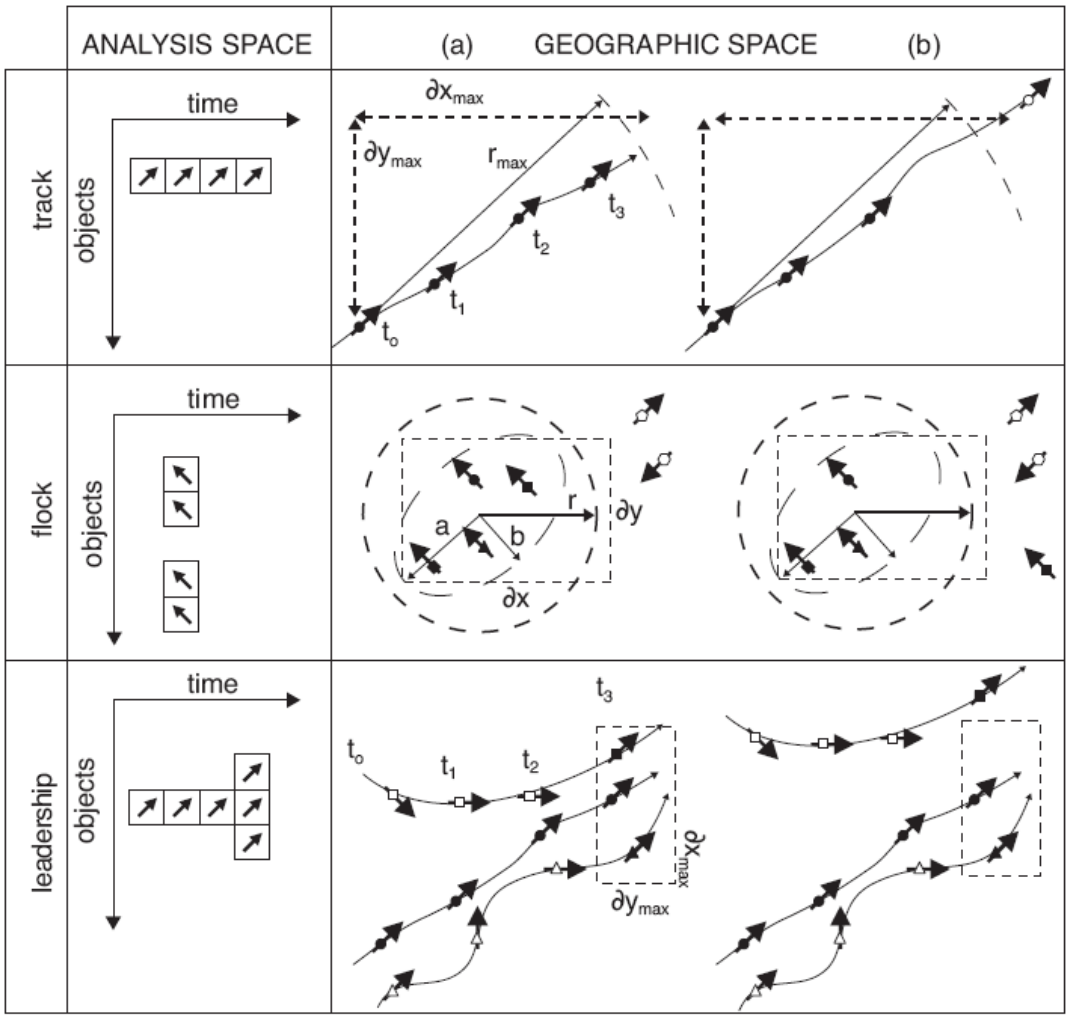

Figure 6: The figure illustrates the constraints of the patterns track, flock and leadership in the analysis space (the REMO matrix) and in the geographic space. Fixes matched in the analysis space are represented as solid forms, fixes not matched as empty forms. Some possible spatial constraints are represented as ranges with dashed lines. Whereas in the situations (a) the spatial constraints for the absolute positions of the fixes are fulfilled they are not in the situations (b): For track the last fix lies beyond the range, for flock and leadership the quadratic object lies outside the range.

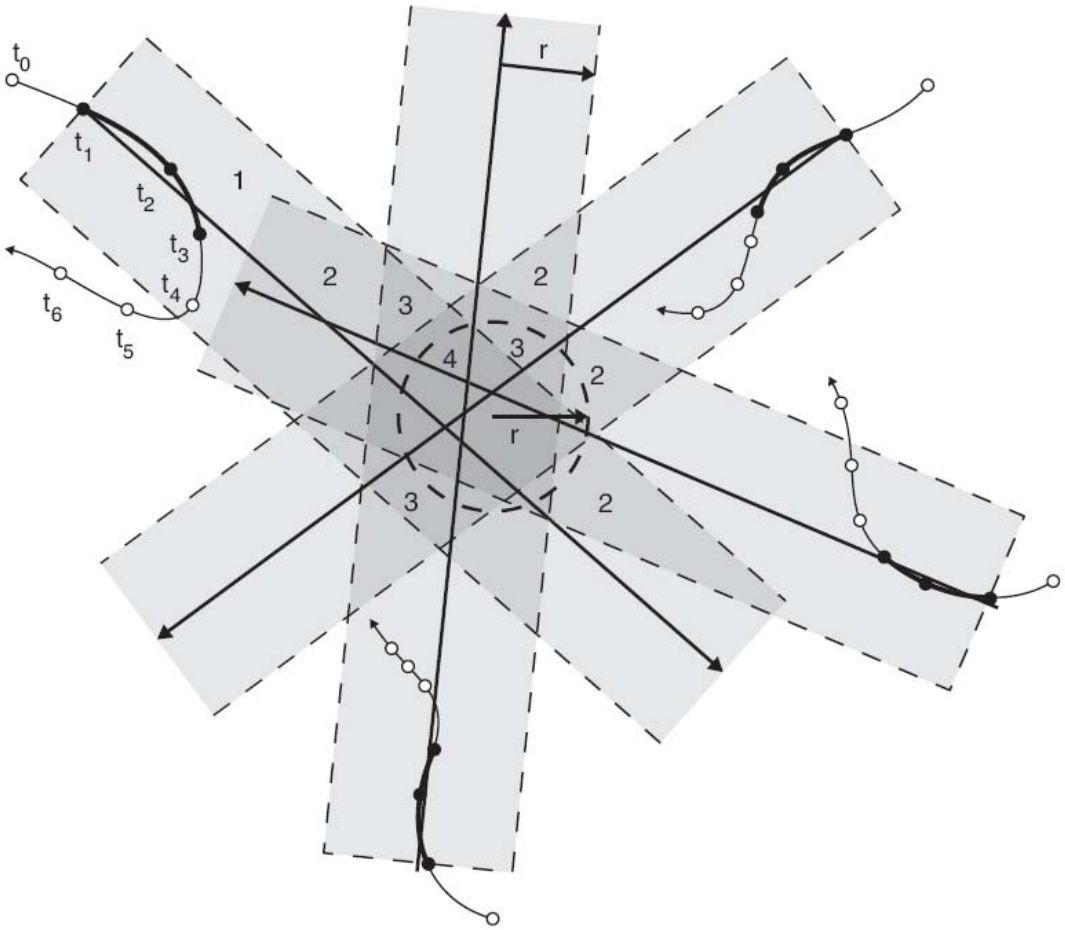

Figure 7: Geometric detection of convergence. Let S be a set of 4 MPOs with 7 fixes from t0 to t6. The illustration shows a convergence pattern found with the parameters 4 MPOs at the temporal interval t1 to t3. The darkest polygon denotes an area where all 4 direction vectors are passing at a distance closer than r. The pattern convergence is found if such a polygon exists. Please note that the MPOs do not build a cluster but nevertheless show a convergence pattern.

2.3.2 Spatio-temporal Association Rules

As illustrated before, most researches until now are studying movement between discretized regions, for example, Verhein et al. (2006) defined spatio-temporal association rules (STARs) that describe how entities move between regions over time. They assume that space is partitioned into regions, which may be of any size and shape. The aim is to find interesting regions and rules that predict how entities will move through the regions. A region is interesting when a large number of entities leaves (sink), a large number of entities enters (source) or a large number of entities enters and leaves (thoroughfare).

A STAR (ri, T_1, q) =) (r_j, T_2) denotes a rule where entities in a region ri satisfying condition q during time interval T_1 will appear in region rj during time interval T_2. The support of a rule is the number, or ratio, of entities that follow the rule. The spatial support takes the size of the involved regions into consideration. That is, a rule with support s involving a small region will have a larger spatial support than a rule with support s involving a larger region. Finally, the confidence of a rule is the conditional probability that the consequent is true given that the antecedent is true. By traversing all the trajectories all possible movements between regions can be modeled as a rule, with a spatial support and confidence. The rules are then combined into longer time intervals and more complicated movement patterns.

Nanni et al.(2008) proposed the case of spatio-temporarily related traffic jams. For example, traffic jam (Pisa, 7.30 AM) ⇒ traffic jam (Lucca, 8.30 AM), meaning that whenever the first event (a traffic jam in Pisa at 7.30 AM) occurs, usually it is followed by the second one (a traffic jam in Lucca at 8.30 AM). A more general version of this rule could be traffic jam (Pisa, t) ⇒ traffic jam (Lucca, t+1 h), in which time appears as a parameter. Rules could also be discovered after even further abstracting time, as the following generalization shows traffic jam (Pisa) ⇒ traffic jam (Lucca). In the same style of these examples, frequent patterns can be discovered in the trajectory data.

Researchers sticking to continuous space and time are trying to tackle with the continuity problem in two complementary ways by (1) considering patterns that are in the form of trajectory segments and searching approximate instances in the data and (2) considering patterns that are in the form of moving regions within time intervals, such as spatiotemporal cylinders or tubes – that, in some sense, represent a segment of trajectory plus a bounded approximation/uncertainty – and counting as occurrences all trajectory segments fully contained in the moving regions.

The work in (Cao et al. 2005) provides an example of the first approach: a trajectory is approximated by means of a sequence of spatial segments obtained through a simplification step and then patterns are extracted essentially in the form of sequences of contiguous spatial segments; in particular, each element of the sequence has to be similar to several segments of the input trajectory, similarity being defined w.r.t. three key parameters: spatial closeness, length and slope angle. Frequent sequences are then

outputted as sequences of rectangles such that their width quantifies the average distance between each segment and the points in the trajectory it covers.

The second approach, based on moving regions, is followed by (Kalnis et al. 2005), and concerns the discovery of density-based spatial clusters that persist along several contiguous time slices. Finally, a similar goal, but focused on cyclic patterns, is pursued in (Mamoulis et al. 2004): the authors define the spatiotemporal periodic pattern mining problem (i.e. finding cyclic sequential patterns of given period) and propose an effective and fast mining algorithm for retrieving maximal periodic patterns. While time is simply assumed to be discrete, spatial locations are discretized dynamically through density-based clustering. Each time a periodic pattern is generated, in the form of a sequence of spatial regions, a check is performed to ensure that all regions in the pattern are dense – and then significant.

2.4 Conclusions

In this chapter, we have applied classification on the input trajectory and output spatio-temporal patterns respectively, and then the patterns and algorithms appeared on papers are discussed. Due to the complexity of these patterns, a clear mapping among the trajectories and patterns is still incomplete. Also, it seems inevitable there may be missed patterns and new patterns will be brought about subsequently. What we are intending to do here is trying to look into the essence of each pattern and find the characteristics in common. So that a novice in this area can quickly gain a systematic view of knowledge that can be extracted from trajectory data and a veteran can easily find inadequate studied areas.

3 Preprocessing and Exploratory Analysis of data

In this chapter, the real data set from Starkey Project will be introduced along with basic background knowledge. Then the detail of data preprocessing will be included in the second section. At last, a first impression of the data set will be gained by exploratory spatio-temporal data analysis, where visualization techniques will be extensively used.

3.1 Starkey data set

3.1.1 Introduction of Starkey project

Starkey Experimental Forest and Range is located 35 km southwest of La Grande, Oregon, in the Blue Mountains of northeastern Oregon, USA. This 10,125-ha project area is enclosed by a 2.4-m high fence that prevents immigration or emigration of resident elk and other large mammals (Rowland et al. 1997). Starkey is divided into multiple subunits, the largest being a 7762-ha main study area where data for the current study were obtained (Figure 8). Starkey is situated at about 1500m elevation and supports a mosaic of coniferous forests, wet meadows and grasslands that typify summer range habitat for elk in the Blue Mountains. A network of drainages creates a complex and varied topography Details of the study area and facilities are available elsewhere (Johnson et al. 2000; Rowland et al. 1997).

Elk locations were obtained by an automated telemetry system that uses retransmitted LORAN-C radio navigation signals (Rowland et al. 1997). A subset of the Starkey telemetry data during the interval of May 2–May 28 in 1996 was selected for this study.

3.1.2 Trajectory data

The data set is available in (US Forest Service 1996). Also available are various explanatory variables describing forest vegetation and topography suspected to influence animal movement. Other habitat features such as distance to road and distance to hiding cover may be derived.

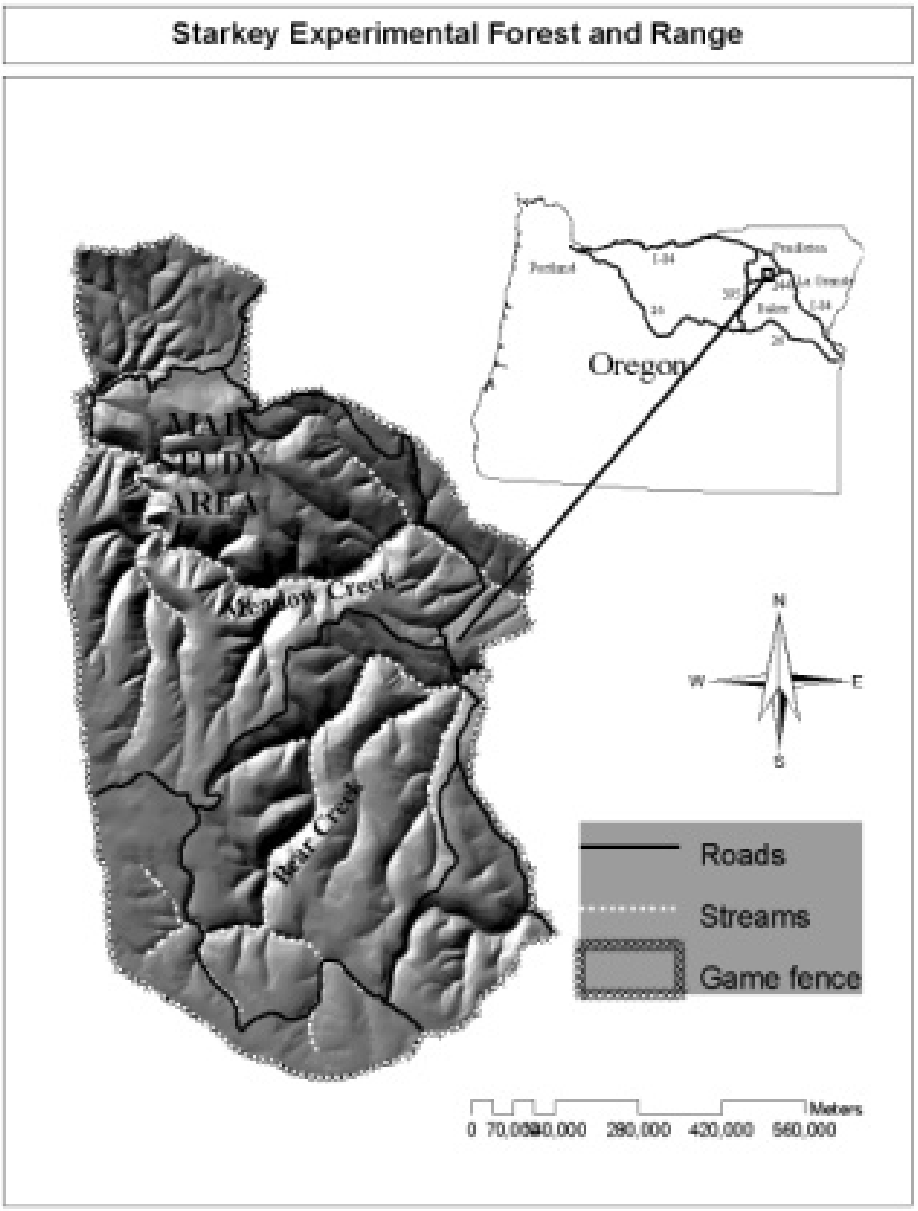

Figure 8: The Starkey Project is located within the Starkey Experimental Forest and Range in the Blue Mountains, northeast Oregon.

The data are spatial-temporal. The animals are labeled by m = 1, ..,M, and their locations are recorded at times, tmk, k = 1, ..,K for the m-th animal. The locations are denoted rm(t), corresponding to the UTM (Universal Transverse Mercator) coordinates of the k-th time measurement of the m-th animal(Brillinger et al. 2004).

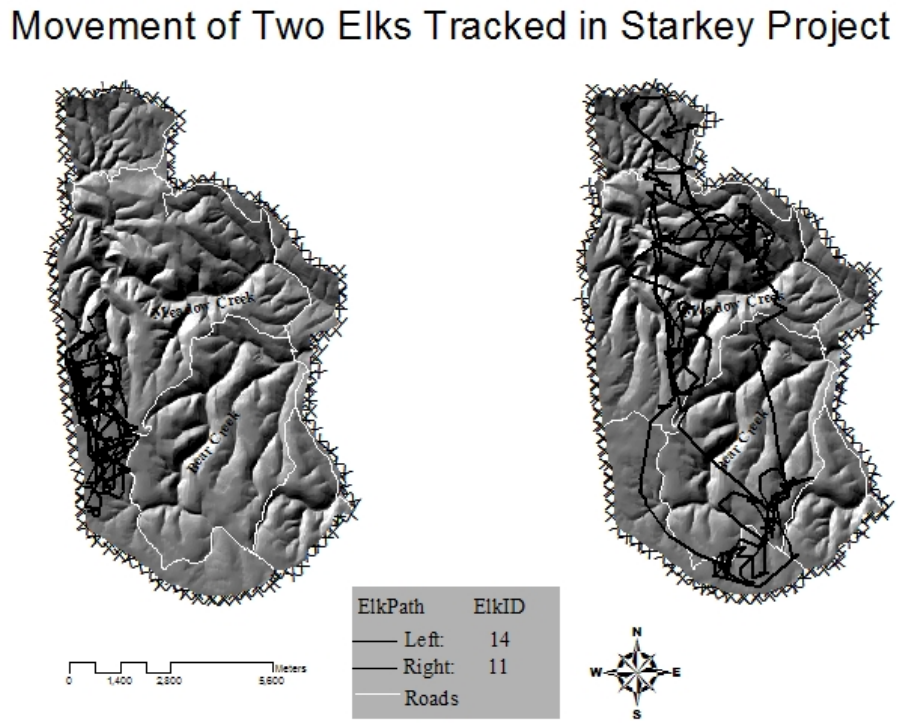

Figure 9: Two elk tracks in May, 1996. The paths are sampled in time, hence the straight line segments.

Figure 9 provides an example of the tracks of two elks. An obvious difference can be observed between these two particular animals: the left one did not move too much over Starkey, but the right one visited widely spread locations all over Starkey.

We used recorded locations of 38 individual elk during 2 May- 28 May (approximately 26 days). The mean elapsed time between locations for each elk averaged 79 min. Locations were assigned habitat information by matching each observation to the closest 30-m×30-m pixel. Locations had a mean error of 53m (Findholt et al. 1996). Calculations of movement were deleted if elapsed time was <5 min or >150 min between successive observations of animals. The Main Study Area at Starkey is 3–4 times larger than typical summer home ranges of elk in the Blue Mountains. This provides elk with large-scale habitat choices commensurate with free-ranging herds. The 38 elks in this study that were tracked simultaneously were single female elk selected at random out of a total population of 311–386 adult cow elk in Starkey.

3.1.3 Previous study on the dataset

There are plenty of studies on this project since 1988 encompassing every aspect such as animal behavior, ecosystem dynamics, distribution, human disturbance, statistical Modeling and so on. The research on movement patterns was mainly conducted by Haiganoush K. Preisler, Alan A. Ager and David R. Brillinger. Several interesting patterns were discovered by EDA and stochastic modeling.

First, a parallel box plot (figure 10) of the square roots of estimated elk speeds by hour of the day. The groups of animals appear substantially more mobile around 0500 hrs and 1800 hrs and less active at night and midday.

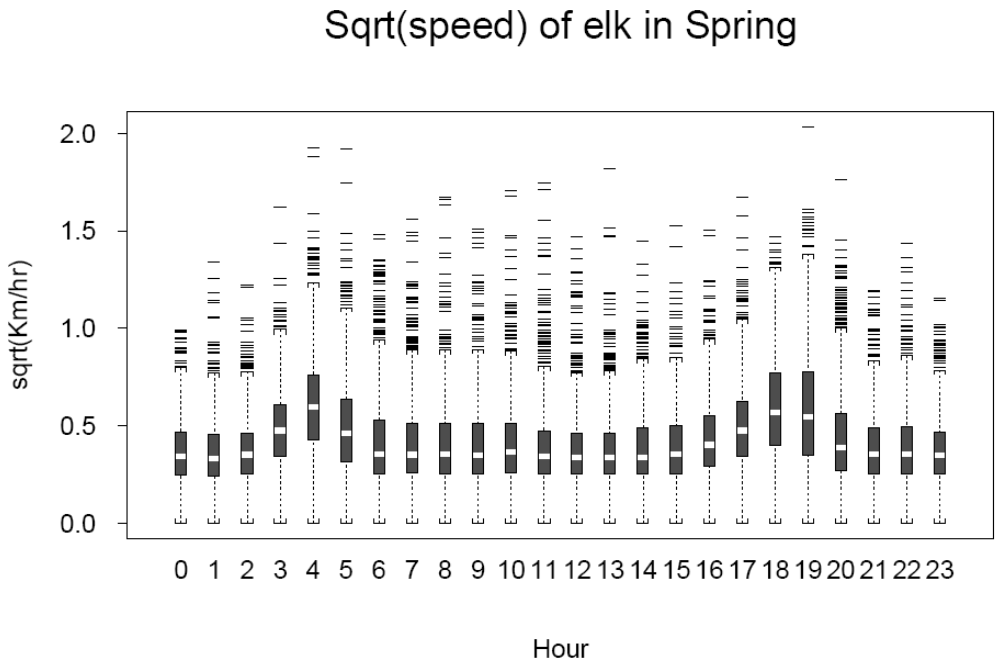

Figure 10: Parallel box plot of estimated elk speeds by hour of the day(Ager et al. 2003)

Secondly, figure 11 provides kernel density estimates of the animals' noon locations based on all the data available. Noon was picked since, following Figure 3, the animals' were less mobile then. There are several hot spots, i.e. locations of congregation, for the elk.

Thirdly, the periodical patterns of environmental variables along elk movement were well studied and reported in detail by Ager et al. (2003). Elk showed pronounced 24-h cycles with crepuscular transitions for many habitat variables, including canopy cover, distance to hiding cover, cosine of aspect, herbage, and distance to open roads. Habitat transitions appeared to be closely linked to rapid changes in elk movements for

most habitat variables, but not all. Morning movements were uphill, towards more convex topography, and at increasing distance to streams. Afternoon movements were directed towards easterly aspects, steeper slopes in valley landforms, and towards streams. At dusk, movements were strongly upslope, out of drainages, and towards areas that are characteristic of foraging areas, i.e. lower canopy cover, greater distance to hiding cover, increased herbage production, closer to roads, and more southerly and westerly aspects.

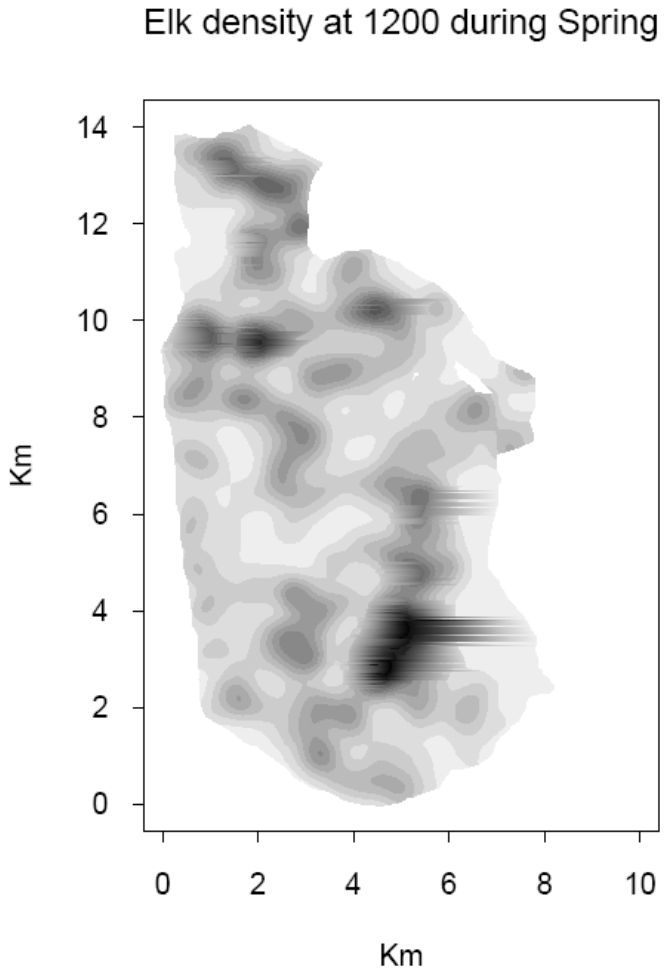

Figure 11: Kernel density estimates of elks at noon during spring(Ager et al. 2003)

Lastly, movement vectors were modeled using bivariate stochastic differential equation, which is comprised of a determinative part and a stochastic part. The former is built from a set of additive potential functions for each habitat variable that affect movement, representing attraction and repulsion to specific habitat features at different times of the day. Four habitat covariates were found that had significant influence on

movement vectors, these being distance to security areas, distance to foraging meadows, distance to steep slopes, and distance to streams. On the other hand, movements that are seemingly random, like foraging paths in a meadow, or movements that cannot otherwise be explained with environmental covariates are included as stochastic terms in the model. The reader is referred to Brillinger et al. (2004) and Preisler et al. (2004) for details.

3.1.4 Elk

The elk (Cervus elaphus), is one of the largest species of deer in the world and one of the largest mammals in North America and eastern Asia. Elk range in forest and forest-edge habitat, feeding on grasses, plants, leaves, and bark.

Some important facts relevant to movement patterns are listed as below(RMEF 1999):

- Food, water, shelter and space are essential to elk survival.
- Female elks, baby elks and yearlings live in loose herds or groups.
- Male elks live in bachelor groups or alone.
- During the rut, female elks and baby elks form harems with one or two mature male elks.
- In cold snowy climates, female elks, baby elks and young male elks migrate to foothills and valleys in winter.
- In spring, male elks begin to migrate first to higher elevations than female elks and young elk. Female elks often begin migrating before they give birth to their baby elks, which are typically born in late May through early June. They stop to give birth and allow their young to grow for several weeks. By late June or July, they've resumed moving into higher country where they will find rich summer food.
- An experienced elk, usually the lead female elk, guides a herd between seasonal ranges.

3.2 Data preprocessing

For many movement descriptors, such as sinuosity, more than one means of computation, including varying parameters such as interval lengths or weight factors. The choice of these naturally has impacts on the final analysis. For example the length of the analytical interval may be as influential as the dimension of a moving window in

other focal operators: an important parameter for analysis but also a major influence as a smoothing filter. This observation is significant in movement research in that often little information is provided about trajectory data models and algorithms employed in computing movement descriptors in a particular study. In order to increase the transparency and the repeatability of analysis of movement trajectories, so how their trajectory descriptors are computed is reported in detail.

3.2.1 Data selection and data cleaning

The main task of data selection is to determine a subset of the records or variables in the database for focusing the search for interesting patterns.

Based on the previous extensive studies on elk trajectories, tracks of 59 elks during spring 1996 were selected at the beginning for two reasons: first the movement of elks was most intensively sampled in 1996, second elks are less active in summer season while requires finer scale study, but both the spatial and temporal accuracy cannot suffice the demands. Furthermore, since the samples were collected temporally uneven, the distribution along monitoring period was studied (Figure 12). As a result, the subset from 2 May to 28 May (shown in the bounding rectangle) of samples was selected.

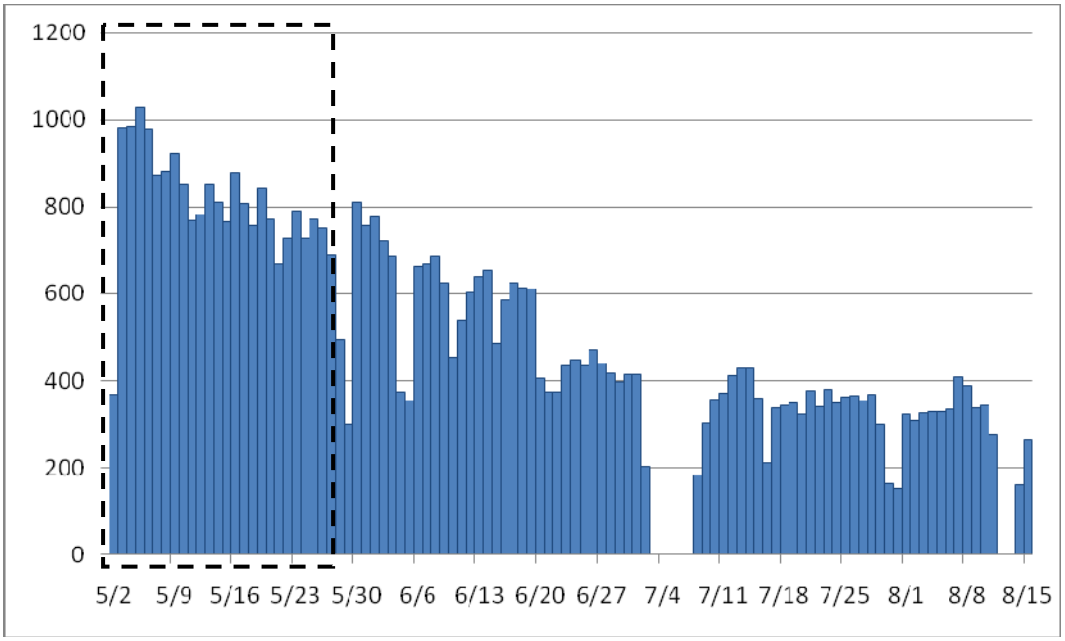

Figure 12: Temporal distribution of samples from 2 May to 15 August, 1996

What was also noticed is the uneven distribution of samples among elks (Figure 13). As to the histogram, elks whose number of collected locations was less than 420 are abandoned due to excessive difference of temporal granularity. This difference can lead to distortion of movement descriptors, for example, lower movement speed and less sinuosity. Due to the reciprocal property of elk movement discovered in previous study, this effect will be more obvious in this case.

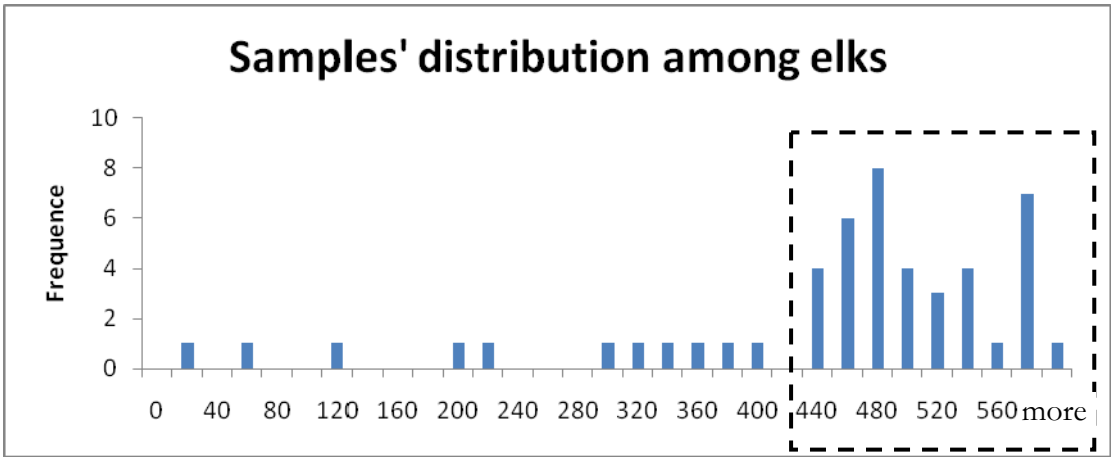

Figure 13: Samples' distribution among elks

In addition, the obvious discorrespondence among the data set such as the contradiction of two timing system (Starkey time and GMT) was observed and relevant samples were eliminated.

At last, due to the unevenly sampled nature, the data set needs to be rediscretized at uniform interval in some cases. Assuming elks are moving along straight line between samples, the locations of elks at every integer point can be interpolated (figure 14).

3.2.2 Data reduction

Data reduction, including transformations, projections and aggregations, is especially important for trajectories to find useful representations. The most fundamental analytical context for tracking data is the single fix, with its most important feature being the location in the embedding geography. Then structures are superimposed on the fixes to comply with that measure's needs, since each parameter, such as speed and acceleration, may require different approaches to aggregate and transform the recorded

fixes. For geospatial lifelines, the lifeline-context operators (Laube et al. 2007) are the conclusion of structural we often used.

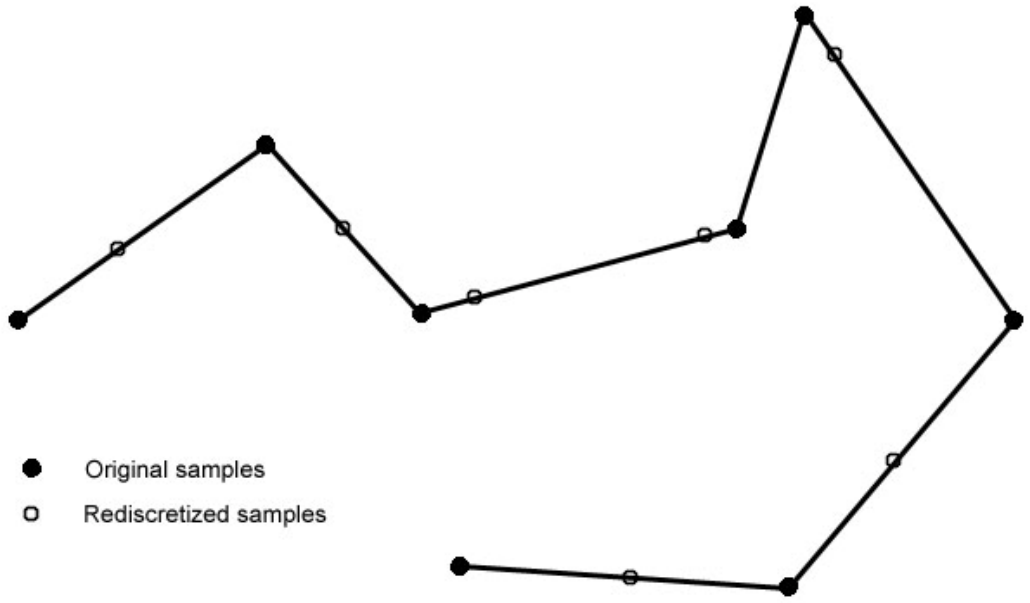

Figure 14: Rediscretization of original unevenly sampled data set

As to Laube et al.(2007) pointed the lifeline-context operators can be classified as instantaneous, interval, episodal, and total in analogy to local, focal, zonal, or global context operators in spatial analysis (figure 15).

- Instantaneous ("local"). Derive $d_t = d(t)$ at an infinitesimal instant in time.
- Interval ("focal"). Use a moving interval (moving temporal window) to investigating a fixed length segment of the lifeline, computing $d_{int} = d(t \pm dt)$ respectively.
- Episodal ("zonal"). Preliminary analysis may result in a partition of the lifeline in delimited episodes, each represented by a movement descriptor $d_{eps} = d[t_{begin}, \ldots, t_{end}]$.
- Total ("global"). Movement descriptors can be computed for whole trajectories as $d_{tot} = d[t_0, \ldots, t_m]$. This is the traditional static perspective of lifelines.

Just as with spatial-context operators, the variability of possible lifeline-context operators seems to be unlimited. Some descriptors of the elk trajectories are computed as follows, while a small part of results are used as an example. However, the speed,

acceleration and some other descriptors should be calculated from original data set instead of rediscretized data set mentioned above.

3.2.2.1 Instantaneous descriptors

Location: as to the rediscretization used before, the hypothetical location of an object at a time can be interpolated from original fixes. Such an interpolation may use a simple interval average or an interval mean of x and y of the involved fixes. In this case, simple interpolation of two neighbor fixes is used.

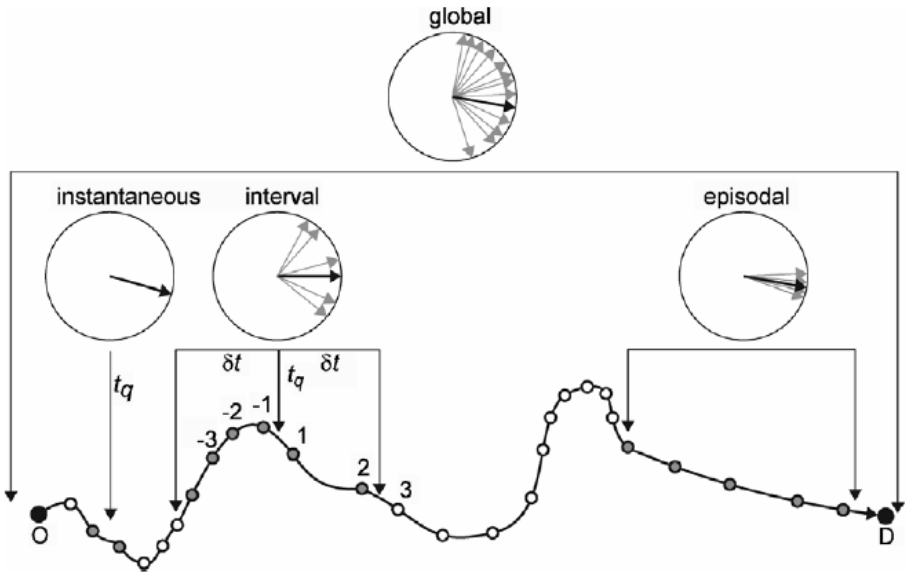

Figure 15: Four different perspectives to derive the movement descriptor azimuth from a lifeline (top). In many cases also instantaneous measures require the inclusion of at least two fixes. Which fixes are included for interval measures may depend on the way the interval is defined. The temporal interval ±dt includes different fixes than using a fixed number of fixes(Laube et al. 2007).

Environmental variable: the environmental variables come along with the locations. In the Starkey project, lots of factors which may influence the movement of elks were monitored, such as canopy, elevation, and distance to road and so on.

Speed and Movement azimuth: The instantaneous velocity of elks at every integer point is computed from original neighbor fixes. This parameter is defined in both magnitude and direction, or magnitude at x and y direction.

Speed at vertical direction: even though the telemetry system did not provide the position of elks at vertical direction, ancillary information can be obtained from digital elevation map. This information is important since elks usually migrate to higher meadow during spring.

There are other instantaneous descriptors such as approaching rate, navigational displacement. However, they are more suitable for trajectories whose starting and ending points are semantically obvious like trajectories of homing pigeons.

3.2.2.2 Interval descriptors

Sinuosity: The terms tortuosity, sinuosity, straightness, and path entropy all refer to the degree of windingness of a trajectory. Laube et al. (2007) listed the most frequent concepts as follows. Tortuosity is normally calculated as the ratio represented by the greatest distance between any two points on the trajectory divided by the length of the path (Benhamou 2004; Claussen et al. 1997). Sinuosity is normally computed from an equal step length rediscretized path, considering the standard deviation of the directional changes and the rediscretization step length (Bovet et al. 1988; Claussen et al. 1997). The straightness index is given as the ratio of the beeline connector distance and the travelled path length (Benhamou 2004; Weimerskirch et al. 2002). Other authors use a fractal dimension of trajectories(Bovet et al. 1988), or an indication for path entropy(Meade et al. 2005; Roberts et al. 2004). Even though these sinuosity measures are mainly applied as total operators, in effect they are perfectly suited for interval and episodal operators. In this case, straightness index is used due to its relatively lower computational intensity. As to figure 3, it takes around 4 hours for elks to reach summit speed and then slow down during dawn and dusk. Combining the influence of coarse temporal resolution, 4 hours' interval is a descent choice.

Interval standard deviation of a descriptor: in certain cases one may want to investigate the distribution of a movement property within a moving interval. For example, higher deviation of movement azimuth often related to foraging activity instead of directional migration. Also, low deviation along with small magnitude of distance to stream indicated that the elk is moving along a stream. The interval standard deviation can also be imposed on interval descriptors like sinuosity, which can be used to find a sudden turn in the trajectory. To compute the deviation, more values are required to ensure the accuracy. On the other side, long interval will to certain extent reduce the responsivity of this descriptor. 6 hours' interval is chosen in this case, however, appropriate temporal resolution is open to discussion.

3.2.2.3 Episodal and total descriptors

Due to the daily cycle discovered in previous study, it's natural to treat the movement of an elk during a day as an episode.

Location: An episodal and total representation of location here refers to the centroid of an episode or the entire trajectory. It indicates that the elk was moving in certain range around the centroid.

Speed: the speed here can be computed using starting and ending point, or using the total distance covered. The former one is directional while the latter one only has magnitude. In this case, as the elks are most inactive during night, the position at 0hr each day can be used to calculate the episodal velocity. On the other hand, the total distance covered everyday can be used to measure the activeness of each elk each day.

Vertical speed: elks usually move to higher elevation for food during spring, so the elevation of foraging area (low canopy cover and high herbage production) is most suitable indicator of migration. An easier solution is to choose the highest value in the whole day.

Area and shape of convex hull: basically, larger convex hulls were produced by more active elks. However, a migrating elk may produce a thin and small convex hull. The shape of convex can be indicated using area divided by the square of perimeter.

3.3 Exploratory Data Analysis

An extensive exploratory Data Analysis was carried out in previous study(Brillinger et al. 2004), thus no more statistical analysis is carried out in this thesis. Two different methods to extend previous study were conceived: first one is zonal statistical analysis by dividing the study area with regular grids. In this way several raster maps can be produced, for example, the value of each raster cell can be how many times this cell has been visited, how many different elks have visited this cell, how long elks has stayed in this cell, what's the average time an elk stay in this cell, how much the average speed is, the deviation of time or speed… It can include all trajectory descriptors if we like. However, no available software can do this kind of work, and to write such software is too time-consuming for this thesis.

Another method we have used is animation. Animated displays are often considered as the first choice when data involve time(Eick 1997). However, psychological studies show that animation is not necessarily effective and superior to static displays(Tversky et al. 2002). It seems that animation is good for gaining an initial overview of a time-related phenomenon or process while the further, more comprehensive exploration requires combination of animation with other displays and rich facilities for user interaction.

4 Knowledge Discovery from Starkey Data Set

In this chapter, the knowledge discovery technique mentioned in chapter 2 will be applied to Starkey data set to find interesting patterns. At first, the trajectories are partitioned using RoIs, which will be derived from the data itself. Then different clustering algorithms will be used to derive meaningful patterns in this case.

4.1 Trajectories partitioning using RoIs

Most trajectories should be partitioned before further analysis due to several reasons. First and most important one is to address trajectories as movements that correspond to semantically meaning full travel instead of only a series of coordinates. For example, a bird that has departed for migration will make a stop somewhere for some time for feeding, another stop for resting and so on until reach the end of its trajectory. Salesperson on business trip will stop at all locations where they planned to meet a customer.

After partitioning, the whole trajectory can be discretized into a sequence of meaningful locations. For migrating birds, it will be a series of habitats on route; for the salesman case, it could be like (home, company A, company B, restaurant A, company C, restaurant B, home). Secondly, similar subtrajectories are more likely to exist than the whole trajectory. For example, a flock of migrating birds may divide into two flocks heading for two habitats after sharing a long trip together. Furthermore, it will be interesting to find similar subtrajectories. For instance, if Kate finds her route from home to school is completely "included" in Jack's way from home to company, then she may ask Jack to share a cab in the morning. Thirdly, due to innate uncertainty, trajectories should be partitioned if the accuracy goes beyond required. For example, two consecutive observation points have 1 day time difference while others are only a few seconds apart.

The third requirement can be easily sufficed by partitioning trajectories at excessively large time interval, while the first and second ones are most met by partitioning using derived Regions of Interest (RoI) from trajectories.

4.1.1 Definition of RoI

Finding RoIs from trajectories has long been an interested subject in LBS (Ashbrook et al. 2003) and Biological behavior study(Carneiro et al. 2008). In addition, most studies treat it as a step in preprocessing to partition trajectories so that semantics can be attached subtrajectories and spatial-temporal sequence pattern can be found(Giannotti et al. 2007).

For various application purpose and data set, several definitions had been proposed in previous study. Porikli (2004) defined a " frequently region" as a minimum bounding rectangle that consists of a set of points, each point of which contains at least MinPts number of neighborhood points in a radius Eps, where MinPts and Eps are user supplied parameters for least number of neighborhood points and neighborhood radius. In database community, "stops" was proposed by Spaccapietra et al.(2008), where a stop is a part of trajectory, such that (1) the user has explicitly defined the part of the trajectory to represent a stop (2) the temporal extent is a non empty time interval and (3) the traveling object does not move (as far as the application view of this trajectory is concerned), i.e. the spatial range of the trajectory for the interval is a single point. Several studies extended this definition in application to find RoI (Alvares et al. 2007; Ashbrook et al. 2003; Palma et al. 2008).

4.1.2 Methodology of detecting RoI

In several contexts the mining problem comes with an a priori knowledge of suitable RoI to apply, manually obtained by experts in the application domain or simply through commonsense. For instance, origin-destination matrices are a common tool for the analysis of urban mobility flows, and both origins and destinations are usually given as background knowledge. In these cases RoI can be conveniently selected among a database of candidate places (e.g., a GIS containing features like restaurants, gyms, shops, etc.) or a subset that satisfy some given criteria (e.g., all shops or all restaurants close to a highway, etc.). However, in some cases, we do not have these information in advance, and therefore they have to be derived someway, as discussed in the rest of this section(Giannotti et al. 2007).

1) Filtration of candidate places

SMOT (Stops and Moves of Trajectories) algorithm (Alvares et al. 2007) was proposed based on the idea of finding stops from candidate places. A stop can be found if the moving object stays long enough in a candidate geographic feature.

2) Point density

The RoIs can be simply derived from point density map, where high density naturally corresponds to high popularity. However, this method suffered several drawbacks as we will see in our case study. Firstly, the threshold value for popularity is arbitrary. Secondly, for different RoIs, the threshold density should be different, for example, the location of city council may have higher density than that of the company you are working for, and the location of the company may have higher density than that of your house. However, if the threshold density is set according to your house, lots of places, like the roads you pass by will be included, which has no semantic meaning in most studies. This disadvantage is especially obvious in animal movement tracks.

3) Direct clustering

Carneiro et al. (2008) tried to extract regions from white storks migration tracks using a combination of Fuzzy c-means, Subtractive and Gaussian Mixture Model clustering technique. Bashir et al.(2005) decompose the whole trajectory into n/P sub-trajectories. Next, all locations from the sub-trajectories which have the same time offset w of P will be grouped into one group Gw. P is data-dependent and has no definite value. For example, P = "a day" in a traffic control application since many vehicles have daily patterns, while animal's annual migration behaviors can be discovered by P = "a year". DBSCAN(Ester et al. 1996) is then applied to find the dense clusters Rw in each Gw. Furthermore, to prevent generating too large clusters, clusters larger than specified threshold will automatically be broken up into two. This method differs from others in considering the dynamic characteristic of RoIs.

4) Two-step method

In most cases, RoIs are often found in two steps: first, dense (i.e. popular) points in space are detected, and then a set of significant regions (i.e. RoIs) are extracted to represent them succinctly.

4.1) Popular places

The detection of popular places is often proprietary. Different data sets provide various opportunities to find meaningful locations.

For tracks of human being, it seems likely that, at least for most people in the city, locations that could be considered significant will be inside buildings where GPS signals do not reach. This means that there will be a stream of recorded data until the user enters a building, then a time gap, and then a resumption of data when the user exits the building. Then Ashbrook and Starner (2003) defined a popular place as any logged GPS coordinate with an interval of time t between it and the previous point. Andrienko et al (2007) used similar method whereas the GPS device is installed on a car and does not work until the car is started.

Giannotti et al.(2007) model the popularity of a point as the number of distinct moving objects that pass close to it w.r.t. a neighborhood function. Then computing the popularity of points is a distinct count problem. If one point is visited by one object several times, one once is counted, so that the popular places found are meaningful for public instead of individual object. In order to reduce the computation complexity, grids are generated in place of computation in continues space.

For a long term trajectory of only one moving entities, Palma et al. (2008) proposed an algorithm named CB-SMOT(Clustering-based Stops and Moves of Trajectories) to find interesting places. The intuition of the method is that the parts of a trajectory in which the speed is lower than in other parts of the same trajectory, correspond to interesting places. In a tourism application, for instance, a tourist's trajectory would be something like: visit an important monument, visit a museum, go to his hotel, go to a night-club, and return to the hotel. Probably his trajectory has a lower speed around these places than it has in other parts of the trajectory where he was moving from one place to another. Then DBSCAN, a famous density-based clustering algorithm is adapted to trajectory data to derive interesting places.

4.2) RoI (Location)

Because multiple location measurements taken in the same physical location can vary, the logger will sure not record exactly the same coordinate for a location even if the user stops for ten minutes at precisely the same point every day. Also, popular places represent extremely fine-grained information that is difficult to handle properly, due to their (typically) large number. For these reasons, clusters of places are often created using a variant of clustering algorithm. The resulting clusters are names locations to make a difference with the term places.

In addition, hierarchical clustering is often used to subdivide a big cluster if there is a network of "sublocations" within it. The sublocations can also be hidden in large scale analysis, e.g. a student's movement inside in a campus will be neglected when analyzing the traffic in the whole city.

4.1.3 Application in Starkey Data Set

The study on Starkey data set encountered several challenge compared with the studies mentioned above. First of all, no specific candidate RoIs available, i.e. the RoIs must be generated from data set. Secondly, compared with human being, animal movements are somewhat random without clear target, especially for grazing activity. Thirdly, the data set itself suffered from several disadvantages, like low temporal resolution, inaccuracy of telemetry coordinates and only a small portion of total population are tracked. In addition, it makes more sense to describe movements relative to the network rather than unconstraint space, because then it is much easier to formulate queries between moving objects and the network(Gueting et al. 2006).

Even though facing so many challenges, an important advantage is the availability of geographic background information, like the slope, canopy, forage production, and so on. Also the previous study on this data set has set up a firm foundation for further exploration.

4.1.3.1 Density distribution and variation with time

At first, consecutive locations with time difference larger than 12 hours are portioned due to the accuracy requirement. Based on the characteristics of the Starkey data set, direct clustering is the most appropriate method for following reasons. (1) No a prior knowledge required; (2) Low temporal resolution excluded the possibility to find stops using threshold value of speed. (3) Unlike tracks of human being and vehicles, no semantics can be derived from large time interval between consecutive points.

Second, the density maps (figure 16) are generated using point density and line density respectively before clustering for exploratory analysis. The difference between the two is line density removed the time effect, i.e. places elks stay longer time are not given higher weight. Thus the results are a little different. For example, in the four oval regions shown on figure16, which basically include two or three dense area, corridors between the dense areas can be observed in line density map instead of point density

map. This may be caused by relatively fast movement between dense regions, thus the points on the corridors are unlikely to be sampled, but line density simply interpolate the points on the corridors and given them equal weights in computing density. Thus the point density is further used for clustering along with coordinates and geographical background information.

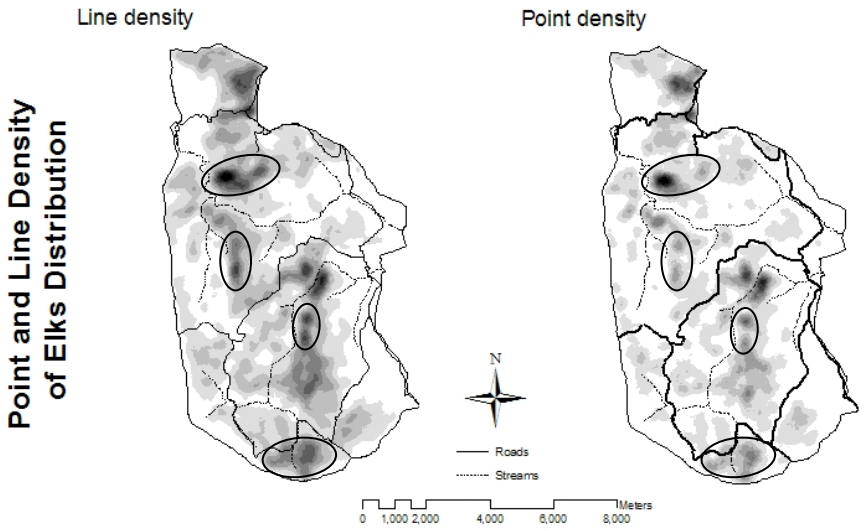

Figure 16: Point and line density of elk distribution

After studying the speed distribution in different hours of a day (figure 10), a closer look of the elk distribution at four different phases, midnight, dawn, daytime and dusk, can be computed and compared (figure 17).

As we can see, during the daytime and midnight, elks are more likely to stay together, i.e. clustered. This provides a great opportunity to improve the result of direct clustering if only the locations obtained during daytime and midnight are considered. Even though low temporal resolution excludes the possibility to find stops using threshold value of speed, these locations during midnight and daytime can be regarded as stops in this application. In addition, two separate clustering during daytime and midnight can identify two different set of RoIs respectively.

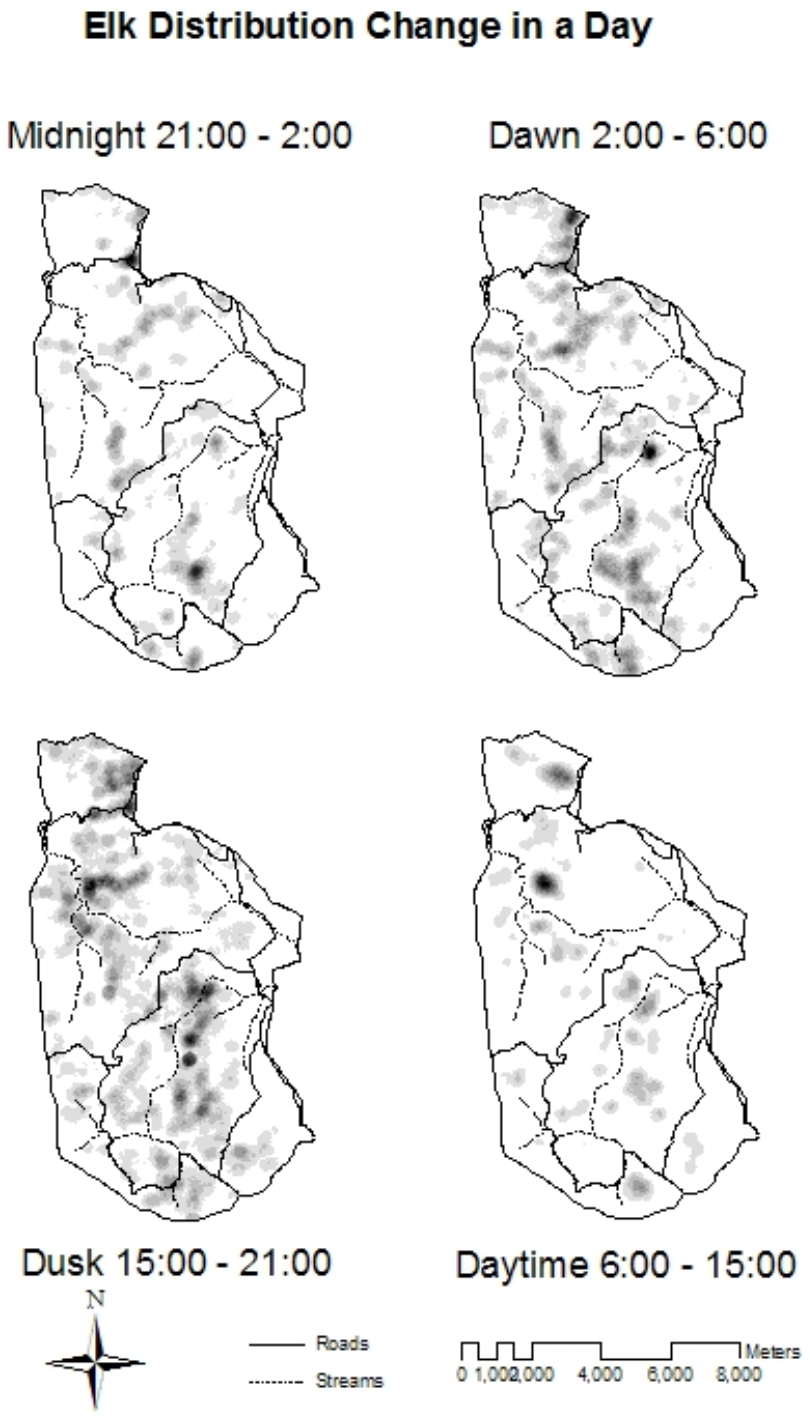

Figure 17: Elk distribution at different period of a day

4.1.3.2 Selection of attributes for clustering

Besides the coordinates and point density, other available relevant geographic information is listed in table 4.

Variable	Variable definition
SoilDpth	Soil depth to the restrictive layer (obtained from the Wallowa-Whitman National Forest soils resource inventory [SRI])
PerSlope	Percent slope; see Rowland et al. (1998)
SINAspct	Sine of aspect; see Rowland et al. (1998)
COSAspct	Cosine of aspect; see Rowland et al. (1998)
Canopy	Total canopy closure (%) of all trees >2.5 cm diameter at breast height)
Elev	Elevation; see Rowland et al. (1998); datum – NAD83; spheroid - World Geographic Reference System 1980
DistEWat	Distance to the nearest water source from within an ungulate-proof pasture (e.g., Main study area), including class I through III streams and water point sources such as stock ponds and springs
DistOPEN	Distance to the nearest open road; see Rowland et al. 1998
DistRSTR	Distance to the nearest restricted access road; see Rowland et al. 1998
DistCLSD	Distance to the nearest closed road; see Rowland et al. (1998)
DistEFnc	Distance to the nearest ungulate-proof fence; see Rowland et al. (1998)
ForgProd	Forage production (biomass of understory species considered forage for ungulates; see Hall 1973)
DistEdge	Distance to nearest edge, based on the EcoGener polygons used for ecoclasses.

Table 4: Potentially helpful environmental variables

Even though so many variables are available, not all of them should be included. For example, the distance to open roads is not important at all during midnight, but plays an important role in daytime. Several statistical and data mining method such as linear regression and decision tree, can be employed to provide insight into the relation between each variable and elk distribution.

In this case a simple linear regression is used to evaluate the importance of each variable. Using these environmental variables as input and density extracted from the previous density maps, the coefficients of each variable at daytime and midnight are listed in table 5.

As we can see from the table, aspect plays the most important role. Specifically speaking, elks are inclined to stay at place with north aspect, especially during midnight, right north aspect is desired. Slope is the second important factor, and elks like to stay at flat plain, especially during daytime. Soil depth and canopy seems to be much less important during midnight compared with daytime. The negative relations during

midnight indicates the preference of less canopy and soil depth, on the other hand, elks likes to stay at places with more canopy and soil depth during daytime.

Quite surprisingly, elevation, distance to water source, distance to roads and forage production are found not so important. However, we can still find some interesting relation. For example, elks mainly stay at lower elevation because the migration has not started for most of them, and this can also be the reason of negative relation with forage production, where high forage production is accompanied with higher elevation in this season; positive relation can be found for distance to all kinds of roads during daytime since obviously elks don't like to stay close to roads while the low importance during midnight is obviously due to the light traffic flow; the negative relation with distance to water source can be explained by the strong influence of slope since the slope are often higher close to water source.

In conclusion, environmental variables marked with grey background along with variables derived from trajectories including coordinates, speed, acceleration, and point density are chosen to be clustered in order to find RoIs. Specifically, attributes such as soil depth, slope, aspect, canopy are included for both midnight and daytime, while an extra attribute, distance to open roads, is included only for daytime.

Attributes	**Midnight Coefficients**	**Daytime Coefficients**
SOILDPTH	-0.46473	3.234282
PERSLOPE	-2.47125	-10.4104
SINASPCT	-27.8966	-12.853
COSASPCT	33.57524	31.88665
CANOPY	-0.33116	2.74133
ELEV	-0.15865	-0.11094
DISTEWAT	0.043	0.082159
DISTOPEN	0.009227	0.308105
DISTRSTR	-0.02634	0.133581
DISTCLSD	-0.02437	0.24145
DISTEFNC	0.000764	0.010678
FORGPROD	-0.00896	-0.19586
DISTEDGE	0.023673	0.205732

Table 5: Coefficients of variables using linear regression

4.1.3.3 Clustering algorithm: GEO-SOM

The clustering algorithm used to cluster the candidate stops is GEO-SOM(Bacao et al. 2004), an adapted Self-Organizing Maps (SOM)(Kohonen 1982) algorithm considering the spatial nature of geographic data. This adaption is done by modifying the way that

Best Matching Units (BMU) are chosen; switching from an equal consideration of all variables, to a more spatially centered one (x and y pairs, for example). That means, units (neurons) that have similar coordinates to the input data are more likely to be considered as BMUs.

To achieve this goal, the search for the BMU is done in two phases(Bacao et al. 2005). In the first phase, only the geographic locations of the data patterns and units are considered, and thus the "first phase BMU" is the unit that is geographically closer to the data pattern being considered. In the second phase, a variable number of units in the output space vicinity of the first phase BMU are considered as candidates to be the final BMU Figure 18. The actual number of units considered in this phase depends on the neighborhood radius t that we have called geographic tolerance. It must be noted that this geographic tolerance is defined in the output space, i.e. in the SOM grid. As a consequence, a given tolerance t corresponds to shorter distances in areas where the geographic density of data is higher, and larger distances where that density is lower. After finding the final BMU the map units are updated according to the standard SOM rule. The choice of t is largely subjective. Because of this different t values should be experimented and the results compared. Basically, t expresses the user's interest in producing local classifications: lower values of t will force the classification of geographic neighboring vectors in closer units.

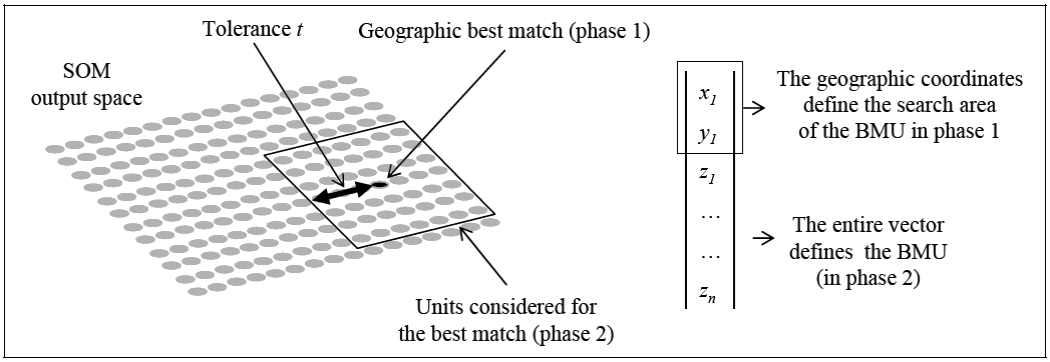

Figure 18: Geo-SOM architecture, showing the unit that is selected amongst all units in phase (1), using only geographical coordinates, and the units that are considered as candidates for BMU in phase (2)(Bacao et al. 2005)

Geo-SOM shows its unique advantage in finding RoIs from candidates in at least two aspects. First, the priority given to spatial position guarantees the spatial closeness of clustering results. Thus the RoIs can be easily derived from each cluster.

second, U-Matrix (Ultsch et al. 1990) and Component Planes provides enough insight into the clustering results compared with other artificial neural network.

Figure 19 shows the clustering results of elk locations during daytime. The figure in the middle is the U-Matrix depicts the distance between units in the output space. Colors varied from deep blue to deep red shows the variance from small to large distance between units. By manually drawn polygons on the U-Matrix, units in the same polygon will be regarded as in one cluster, i.e. the distance between these units is much smaller than that between these units and other units outside the polygon. To make sure of this, the borders of these polygons usually lie on brighter hexagons. The component plain of DTDENSITY (density during daytime) and speed are shown on the right along with the input points on the left. As we can see, not all locations are assigned a cluster; this is because our purpose is only to find the popular locations, i.e. with relatively high density. Thus special attention is paid on the unit with higher density. Actually, visually detected clusters from U-Matrix show a great correspondence with the higher density area in output space, like the first cluster denoted by bubbles on four maps separately showing the densest area on the map. We can also find that all clusters detected except the smallest one are corresponding to low speed area on the speed component plane.

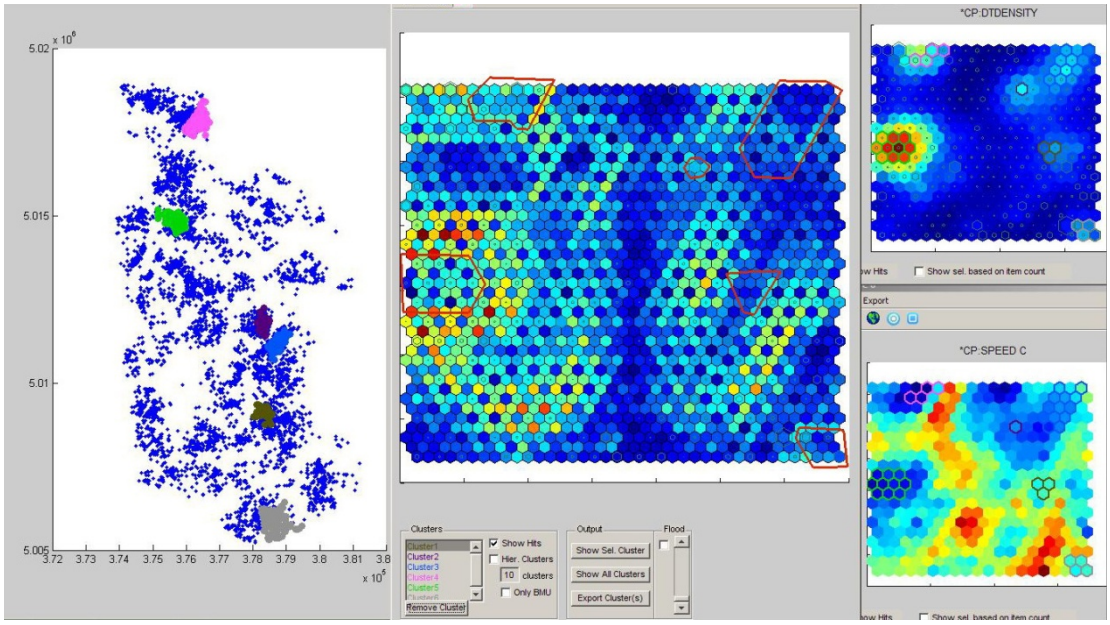

Figure 19: Clusters (left), U-Matrix (center) and Component plane of density (upper right) and speed (lower right) after applying Geo-SOM on elks' locations during 6:00 to 15:00 (t=2)

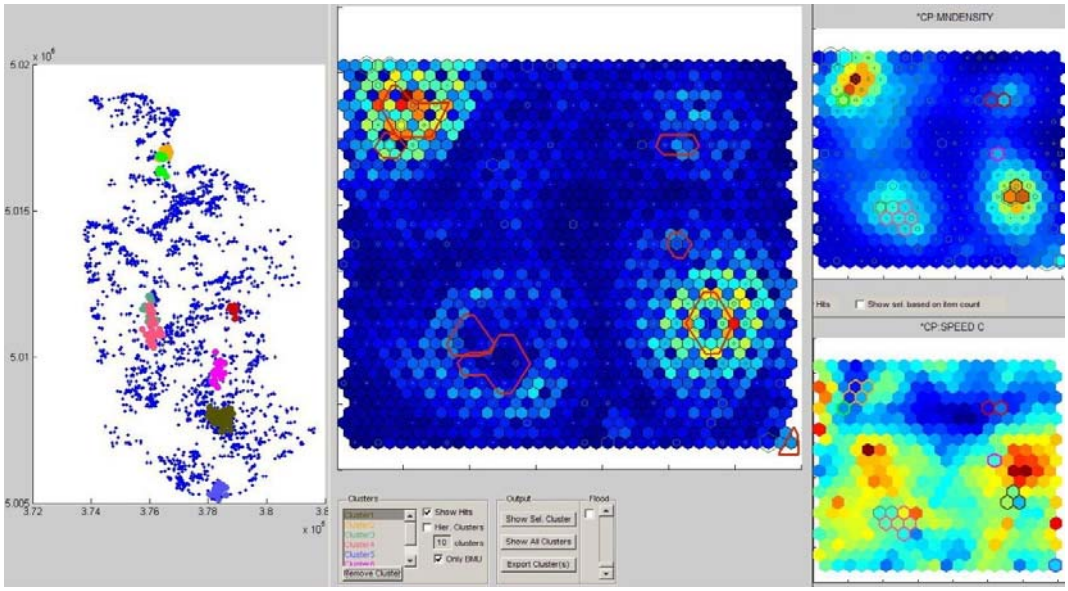

Figure 20: Clusters (left), U-Matrix (center) and Component plane of density (upper right) and speed (lower right) after applying Geo-SOM on elks' locations during 21:00 to 2:00 (t=2)

Figure 20 shows similar clustering results of elk locations during midnight. The only difference is two more clusters are detected from the U-Matrix. The resulting two pairs of polygons are so close to each other that the locations in different clusters are overlapping in space. This can be solved by choosing a smaller spatial tolerance. In this case, we can easily modify the shape of RoIs.

4.1.3.4 RoIs generation and trajectory partitioning

Based on the clusters of locations generated using Geo-SOM, the generation of RoIs usually has several options such as minimum or buffered convex hulls, bounding circles or bounding rectangles. In this case minimum convex hull is used and the RoIs are generated as shown in figure 21.

The result shows overlapping between the RoIs during daytime and midnight. But this overlapping is only spatial while they are temporally disjoint. The trajectories crossing daytime RoIs at night will not be partitioned consequently.

4.2 Trajectory clustering

4.2.1 Trajectory partitioning

After partitioning using RoIs derived above, the trajectories produced by 38 elks are divided into 1086 subtrajectories when the minimum time to stay in the RoIs is 300 seconds. To make sure the trajectories to be clustered are of similar size, a histogram of

duration of each trajectory is generated (figure 22). An extreme case is observed, that is elk 14 has never been to the RoIs at all. After taking a closer look at the histogram in the dense area (figure 23), we can observe that most trajectories produced have duration less than 24 hours. After studying the long duration trajectories, we find that reasons why they are not divided can be concluded as follows:

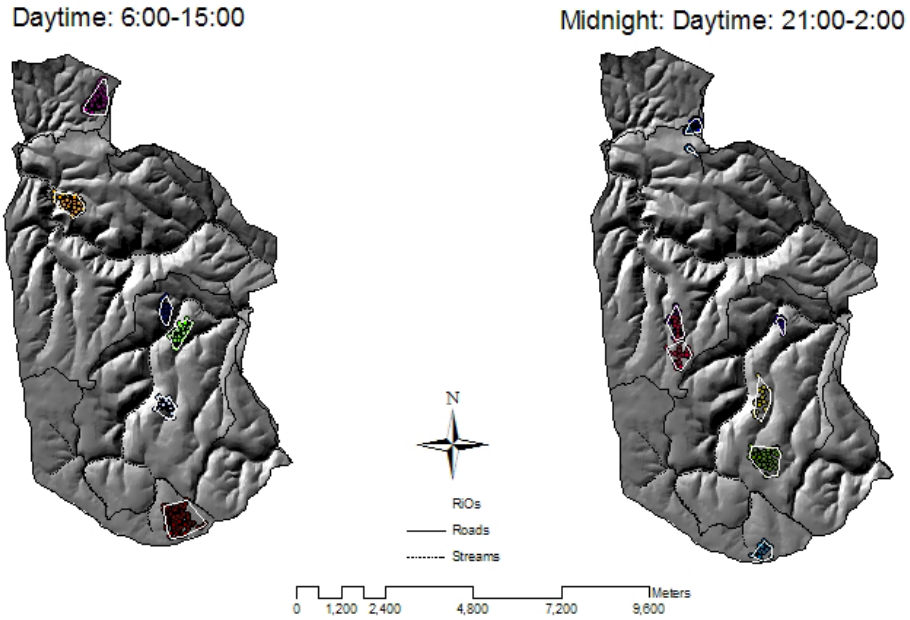

Figure 21: RoIs produced by Minimum Convex Hull

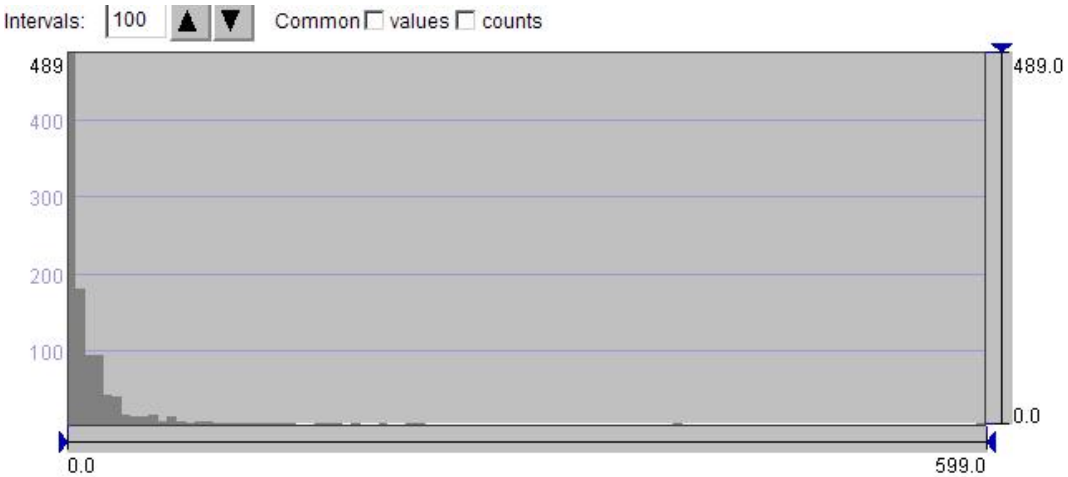

Figure 22: Histogram of trajectory durations after partitioning

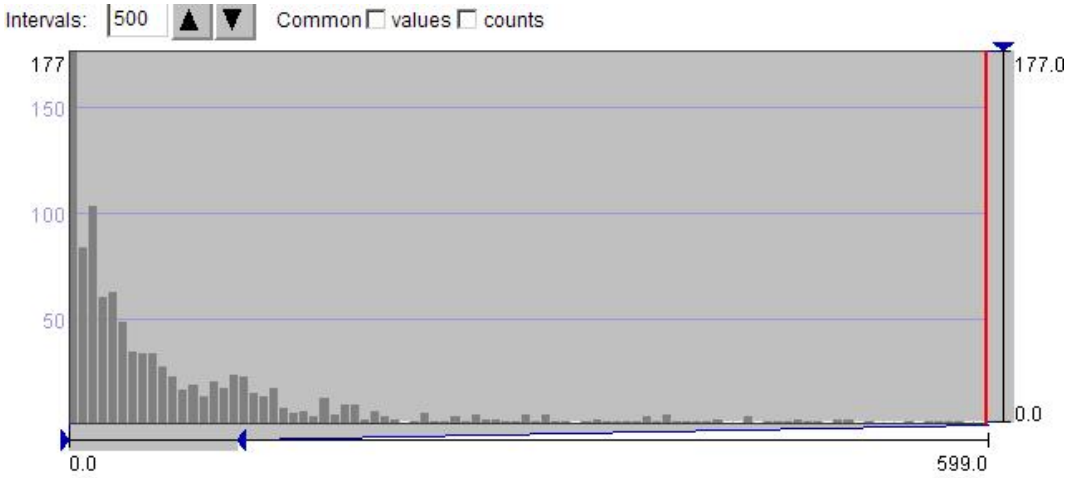

Figure 23: Histogram of the least 20% of trajectory durations after partitioning

1) The low accuracy of telemetry data. Some trajectories are often only a few meters away from the RoIs, thus a buffer of the convex hull can help solve problem of this kind.

2) There are other RoIs unidentified. First only 38 elks are tracked which is only a small portion of the whole elk population, so that some frequently visited area may not be able to be found.

3) Special trajectory of some elks: take elk 14 as an example, which has never been to any RoIs, it keeps staying at the high elevation area of the whole range. This may be because it migrates earlier than other elks, or because she is pregnant as some study shows pregnant elk often give birth to her children at remote location.

Another problem we find is lots of trajectories with too short duration, and this is because of too short duration threshold setting for partitioning the trajectories.

So at last, a buffer of 200m is applied to the RoIs produced by minimum convex hulls, and the duration threshold set for defining a stop at the trajectory is set to 1 hour, as a result, the trajectories are divided into 887 subtrajectories. Then 682 of all the trajectories whose duration is between 2 hr and 48 hr are used clustering,

4.2.2 Trajectory clustering

As we have said before, perform clustering on trajectories can have various patterns to be retrieved depending what we want. In this case, our intension is to find if there are some corridors, i.e. frequently used route, between elks daytime and night RoIs. As to our classification, what we want is clusters considering massive controlling points along

the trajectories while ignore temporal extension. Thus the distance is defined using the built-in function of route similarity in CommonGIS(Andrienko et al. 2007).

A multiple step clustering method is applied in this case. First two distinct clusters are generated (figure 24). Cluster 1 in red is mainly showing the corridors on the upper part, while cluster in blue shows the movement in the lower part. Two frequently used routes are identified using transparent yellow polygons, which can be regarded as corridors between the daytime and night RoIs. As we can see, there are few trajectories that go between these two distinct parts, which is obviously the result of big canyon between them.

A second clustering is then applied to cluster 2, which results in three new clusters shown in figure 25. Cluster 1 is mainly trajectories moving around RoI N1 and cluster 2 is the movement between RoI N1 and RoI N2. Also, rare trajectories between RoI N1, RoI N2 and cluster3 are found, which may be attributed to the road between them.

A third clustering is now applied to the cluster 3 produced in the last step and the result is shown in figure 26. Similarly, two new clusters are produced. Cluster 1 in blue shows the movement between the lowest two RoIs and again little connection between the two clusters can be found, again, it maybe ascribe to the road between them.

After the fourth and the last clustering, another two distinct clusters are shown in figure 27. To summarizing all the clusters, we have identified 6 clusters, of which two clusters are around only night RoIs (clusters in blue and red in figure 25). It can be referred that there may be a daytime RoI nearby which was not found before. The corridors we found only exist in the upper part (figure 24). Connection between other RoIs is either obstructed (by canyon and roads) or cannot be identified because the RoIs are too close to each other. The line density map (figure 16) may help in the later case.

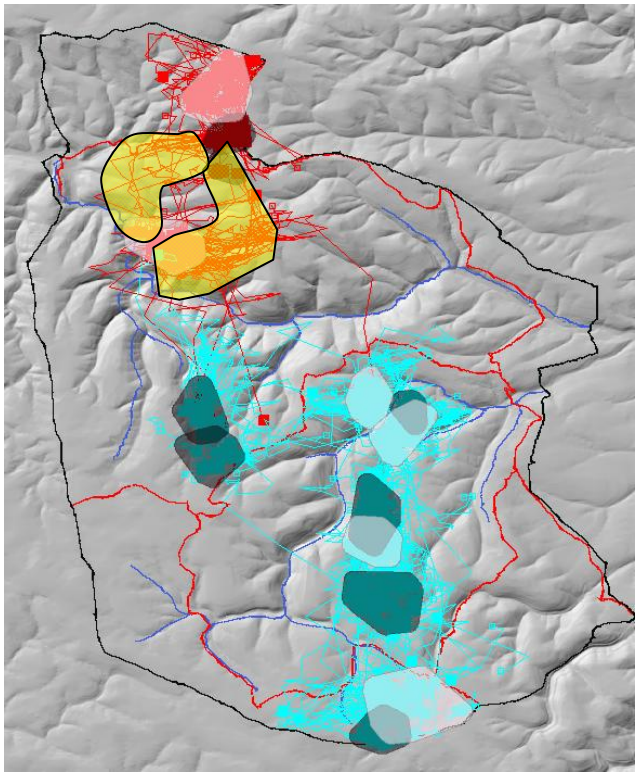

Figure 24: Two distinct clusters generated at first. Polygons in white shade are buffered RiOs during daytime while polygons in dark shade are buffered RiOs during the night.

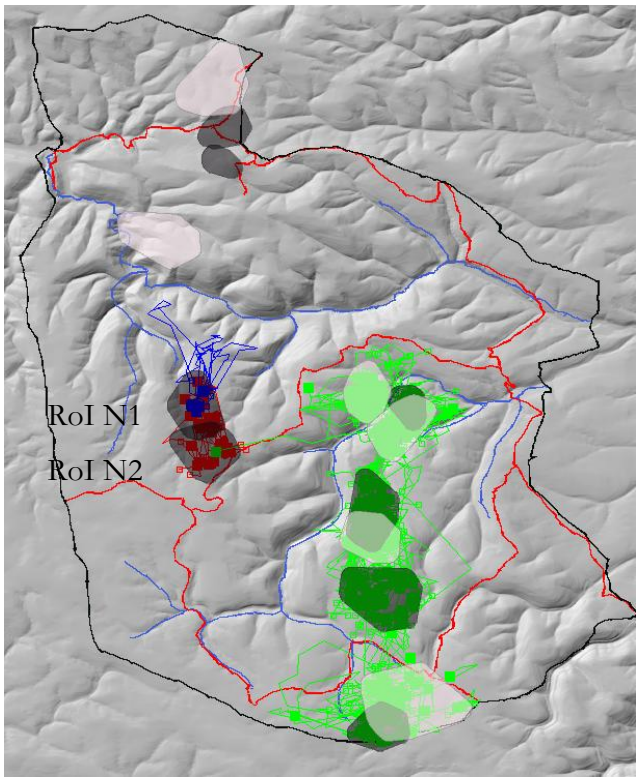

Figure 25: Three clusters generated at the second step.

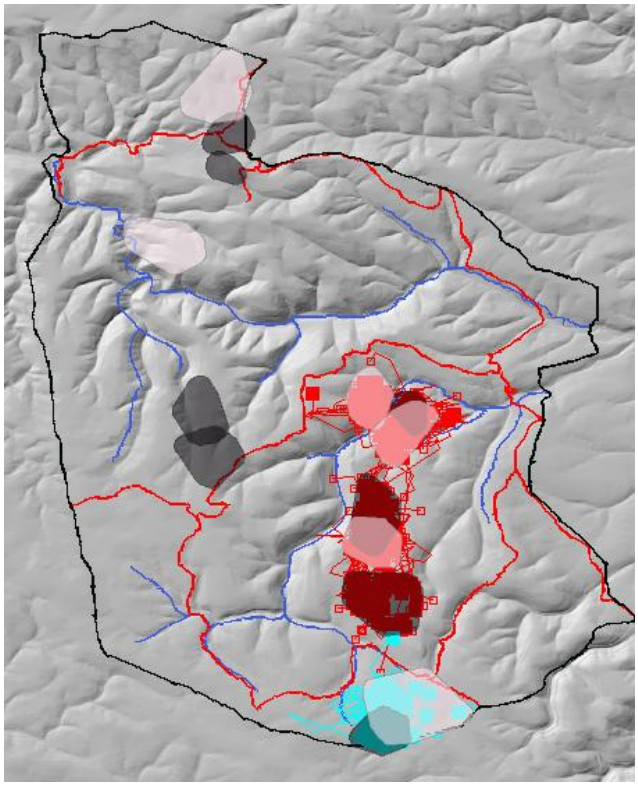

Figure 26: Three clusters generated after third clustering.

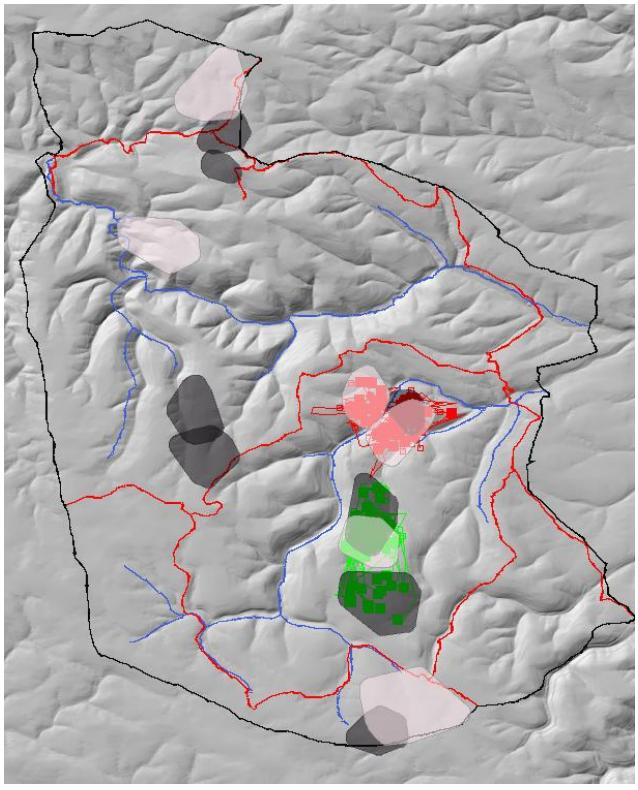

Figure 27: Two clusters generated after fourth clustering.

4.3 Conclusions

In this chapter, knowledge discovery processes were applied to the Starkey data set. Based on previous study, we identified two periods when elks are moving at lower speed and more centralized at some regions. By applying clustering algorithm on the locations of elks at these two periods respectively, we found RoIs at daytime and night. Then trajectories are partitioned using the RoIs detected. After filtering too short or too long trajectories, hierarchical clustering is then applied to find corridors between these RoIs. This clustering algorithm considers massive controlling point along the route, while discard temporal information completely. The result shows two corridors visually detectable (figure 24) and in most cases, connection between other RoIs is either obstructed (by canyon and roads) or cannot be identified because the RoIs are too close to each other.

5 Conclusions and Future Research

5.1 Conclusions

This thesis explored the complexity of spatio-temporal pattern finding from both theoretical and practical aspect. In the theoretical exploration, a framework for spatio-temporal patterns is built and the traditional patterns are put into corresponding class in the taxonomy. In the practical part, a real life data set is used to show the methodology used for knowledge discovery from trajectories. It also shows the effectiveness brought by a clear classification of spatio-temporal patterns. Based on these experiences, we concluded:

a) The complexity of spatio-temporal patterns is brought by the special features of spatial and temporal variables. For spatial attributes, such as coordinate pairs or a cell in grids, implicit topological relations and autocorrelation constitute an important source of the complexity of spatio-temporal patterns. For time, which is innate directional and cyclical, also have similar topological relations and autocorrelation. In addition, both spatial and temporal attributes can be treated at different measurement scales: ratio, interval, ordinal or even nominal. Lastly, last dimension, ID, is also a potential variable.

b) Basically, spatio-temporal patterns can be classified into two major classes: clusters (including classes, clusters, and outliers) and spatio-temporal association rules.

c) For clusters, the most basic form of cluster is flocking, i.e. moving objects keep moving together at every moment. However, this kind of pattern is often too restrictive in our study, so it can be loosen in three dimensions. First, in ID dimension, a new pattern named moving clusters is brought about if we treat ID as a variable instead of constant. Second, in spatial dimension, if coordinates are regarded as interval attribute, spatial translation and rotation will be allowed, as a result, trajectories of similar shape will be clustered. If coordinates are further downgraded to ordinal, which is quite rare, spatial distortion will be introduced in. Third, in temporal dimension, temporal translation will be allowed if time is thought as an interval attribute, resulting clusters of trajectories produced by objects moving at similar route with similar pace but at different time. If time is ordinal, the restriction on pace will be removed. Another

unique feature, cycling, of time can be added and lead to ratio-cyclic, interval-cyclic or ordinal-cyclic patterns.

d) The partitioning of trajectories can have great influence on clustering results depending on how and how many controlling points are compared. We refer the basic form of clusters mentioned above as having massive controlling points along trajectories, then how these points are selected can play an important role. For example, for two trajectories partly overlapping, besides applying partitioning beforehand, we can choose controlling points at certain distance or time interval, and then try to look for the longest consecutive similar points or at longest time interval. An important simplifying technology is to reduce the number of controlling points. Trajectories can be reduced into a series of landmark points or just starting and ending points or even only one point, for example, clustering all trajectories that has visited Lisbon.

e) The basic form of clusters can be extended in ID, spatial and temporal dimension to form complex clustering patterns. In addition, some interesting patterns can be regarded as combinations of clustering problem. Take leadership as an example, leader of a flock can be detected by using clustering algorithm twice. First time, a clustering allowing temporal translation is applied to the trajectories, thus clusters of moving objects following similar routes are generated. Second time, clustering on time dimension is applied to each resulting clusters from first step. Then the leader can be regarded as an outlier temporally ahead of each flock, which is the cluster produced in second step.

f) For spatial-temporal association rules, most studies are using discretized regions in analogy to items in shopping basket. First, in temporal dimension, if time is ordinal, the association rule mining problems become FSP (Frequent Sequential Patten) problem which can be solved using some traditional algorithms. If time is interval, the appearing time in these regions will be have fixed time interval, furthermore, if time is ratio, the appearing time in these regions will be at some fix time. Another extension is to use salient features instead of regions, for example, a landmark location, a U-turn, a sharp increase or decrease of speed and even a piece of subtrajectory. Last extension is to take spatial topology into account, especially to consider the regions is touched or disjoint.

g) For spatial-temporal association rules, sometimes we are looking for the rules, while we also often look for the time and space when and where these rules occur. This class of association rule mining is referred as occurrence retrieval.

h) In the practical part, some knowledge discovery processes are applied to the Starkey data set. Based on previous study, we identified two periods when elks are more gathered and less moving. By applying clustering algorithm on the locations of elks at these two periods respectively, we found RoIs at daytime and night. Then trajectories are partitioned using the RoIs detected. After filtering too short or too long trajectories, hierarchical clustering algorithm is applied to find corridors between these RoIs. This clustering algorithm considers massive controlling point along the route, while discard temporal information completely. The result shows two corridors visually detectable and in most cases, connection between other RoIs is either obstructed (by canyon and roads) or cannot be identified because the RoIs are too close to each other.

In conclusion, a systematic and scientific taxonomy of spatio-temporal patterns will greatly facilitate the knowledge discovery process by quickly identify the tasks for data mining and choose appropriate methodologies. It can further help researchers both in looking for research opportunities and communication and in this field

5.2 Future research

As a newly booming research area, there are plenty of possibilities need to be further studied.

1) The taxonomy proposed in this thesis is still incomplete. In the future, this system need further improvement and keep incorporating newly proposed patterns.

2) A content management system can be built on this taxonomy to facilitate researchers looking for relevant publications.

3) Knowledge discovery software for trajectories can be built on this system. Until now, most algorithms available are proprietary and applied using all kinds of programming platforms. GI software tools incorporation the capability of preprocessing, spatio-temporal data mining, visualization will be a great help for the whole community.

4) The algorithms missed to bridge input trajectories and output patterns are still in need of development. For example, to cluster similar trajectories with extraordinary

noise, a voting algorithm can be a good solution. Another example is the movement pattern of animals scaring away by a moving car, which is a more complex form of divergence.

5) Preprocessing software for trajectories is an emergent need. This software should support drill down and roll up in ID, spatial and temporal dimensions at all kinds of user-defined level of aggregations. Especially at spatial dimension, the spatial analysis capability in GI system should be fully utilized.

6) Even though the data set from Starkey project is quite old and subsequently has low temporal resolution and high uncertainty, interesting patterns still can be found with a creative mind.

BIBLIOGRAPHIC REFERENCES

Ager, A. A., et al. (2003). Daily and seasonal movements and habitat use by female rocky mountain elk and mule deer. Journal of Mammalogy **84**(3): 1076-1088.

Agrawal, R., et al. (1995). Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. Proceedings of the International Conference on Very Large Data Bases, Institute of Electrical and Electronics Engineers (IEEE): 490-501.

Agrawal, R. and R. Srikant (1994). Fast algorithms for mining association rules. Proc. 20th Int. Conf. Very Large Data Bases, VLDB: 487-499.

Alvares, L. O., et al. (2007). A model for enriching trajectories with semantic geographical information. Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems. Seattle, Washington, ACM**:** 162-169.

Andersson, M., et al. (2007). Reporting leadership patterns among trajectories. Proceedings of the 22nd ACM Symposium on Applied Computing, ACM: 393-397.

Andrienko, G., N. Andrienko and S. Wrobel (2007). Visual analytics tools for analysis of movement data. SIGKDD Explorations Newsletter **9**(2): 38-46.

Andrienko, G., et al. (2006). Mining spatio-temporal data. Journal of Intelligent Information Systems **27**(3): 187-190.

Anselin, L. (1998). Exploratory spatial data analysis in a geocomputational environment. Geocomputation: a primer. New York, John Wiley & Sons**:** 77-94.

Ashbrook, D. and T. Starner (2003). Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing **7**(5): 275-286.

Bacao, F., V. Lobo and M. Painho (2004). Geo-Self-Organizing Map (Geo-SOM) for Building and Exploring Homogeneous Regions. Lecture Notes in Computer Science: 22-37.

Bacao, F., V. Lobo and M. Painho (2005). Geo-SOM and its integration with geographic information systems. 5th Workshop On Self-Organizing Maps, Paris.

Bashir, F. I., A. A. Khokhar and D. Schonfeld (2005). Object Trajectory-Based Activity Classification and Recognition using Hidden Markov Models. IEEE Trans. Image Process **16**(7): 1912-1919.

Batty, M., J. Desyllas and E. Duxbury (2003). The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. International Journal of Geographical Information Science **17**(7): 673-697.

Benhamou, S. (2004). How to reliably estimate the tortuosity of an animal's path: straightness, sinuosity, or fractal dimension? Journal of Theoretical Biology **229**(2): 209-220.

Benkert, M., et al. (2007). Reporting flock patterns. Computational Geometry: Theory and Applications **41**(3): 111-125.

Berndt, D. and J. Clifford (1994). Using dynamic time warping to find patterns in time series. KDD Workshop 1994: 359-370.

Bovet, P. and S. Benhamou (1988). Spatial analysis of animals' movements using a correlated random walk model. Journal of Theoretical Biology **131**(4): 419-433.

Bozkaya, T., N. Yazdani and M. Özsoyoğlu (1997). Matching and indexing sequences of different lengths. Proceedings of the sixth international conference on Information and knowledge management, Las Vegas, Nevada, USA, ACM New York, NY, USA: 128-135.

Brillinger, D. R., et al. (2004). An exploratory data analysis (EDA) of the paths of moving animals. Journal of Statistical Planning and Inference **122**(1-2): 43-63.

Cao, H., N. Mamoulis and D. W. Cheung (2005). Mining frequent spatio-temporal sequential patterns. Proceedings of the Fifth IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA: 82-89.

Carneiro, C., et al. (2008). Advanced Data Mining Method for Discovering Regions and Trajectories of Moving Objects: "Ciconia Ciconia" Scenario. The European Information Society: Taking Geoinformation Science One Step Further, Springer Berlin Heidelberg**:** 201-224.

Cheng, C., R. Jain and E. v. d. Berg (2003). Location prediction algorithms for mobile wireless systems. Wireless internet handbook: technologies, standards, and application, CRC Press, Inc.**:** 245-263.

Chrisman, N. R. (2002). Exploring geographic information systems, 2nd Edition, Wiley New York.

Chudova, D., et al. (2003). Translation-invariant mixture models for curve clustering. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., ACM New York, NY, USA: 79-88.

Claussen, D. L., M. S. Finkler and M. M. Smith (1997). Erratum: Thread trailing of turtles: methods for evaluating spatial movements and pathway structure. Canadian Journal of Zoology **75**(12): 2120-2128.

Dobson, J. E. and P. F. Fisher (2003). Geoslavery. IEEE Technology and Society Magazine **22**(1): 47-52.

Dumont, B., et al. (2005). Consistency of animal order in spontaneous group movements allows the measurement of leadership in a group of grazing heifers. Applied Animal Behaviour Science **95**(1-2): 55-66.

Eick, S. (1997). Engineering perceptually affective visualizations for abstract data. Scientific Visualization Overviews, Methodologies and Techniques, IEEE Computer Science Press**:** 191–210.

Ester, M., et al. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press: 226-231.

Fayyad, U. M., G. Piatetsky-Shapiro and P. Smyth (1996). From data mining to knowledge discovery: an overview. Advances in knowledge discovery and data mining. Menlo Park, CA, USA, American Association for Artificial Intelligence**:** 1-34.

Findholt, S. L., et al. (1996). Corrections for position bias of a LORAN-C radio-telemetry system using DGPS. Northwest Science **70**(3): 273-280.

Gaffney, S. and P. Smyth (1999). Trajectory clustering with mixtures of regression models. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, United States, ACM: 63-72.

Ganskopp, D. (2001). Manipulating cattle distribution with salt and water in large arid-land pastures: a GPS/GIS assessment. Applied Animal Behaviour Science **73**(4): 251-262.

Giannotti, F., et al. (2007). Trajectory pattern mining. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA, ACM: 330-339.

Gudmundsson, J., P. Laube and T. Wolle (2008). Movement patterns in spatio-temporal data. Encyclopedia of GIS. S. Shekhar and H. Xiong, Springer, Berlin.

Gueting, R. H., V. T. de Almeida and Z. Ding (2006). Modeling and querying moving objects in networks. The VLDB Journal The International Journal on Very Large Data Bases **15**(2): 165-190.

Hägerstraand, T. (1970). What about people in regional science? Papers in Regional Science **24**(1): 7-24.

Hägerstrand, T. (1976). The time-space trajectory model and its use in the evaluation of systems of transportation. International Conference on Transportation Research, Vienna.

Han, J., et al. (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery **8**(1): 53-87.

Hwang, S. Y., et al. (2005). Mining mobile group patterns: A trajectory-based approach. Advances in Knowledge Discovery and Data Mining, Springer Berlin**:** 713-718.

Johnson, B. K., et al. (2000). Resource selection and spatial separation of mule deer and elk during spring. Journal of Wildlife Management **64**(3): 685-697.

Kalnis, P., N. Mamoulis and S. Bakiras (2005). On Discovering Moving Clusters in Spatio-temporal Data. Lecture notes in computer science **3633**: 364.

Kohonen, T. (1982). Clustering, taxonomy, and topological maps of patterns. Proceedings of the 6th International Conference on Pattern Recognition, Munich: 114-128.

Kritzler, M., M. Raubal and A. Kruger (2007). A GIS Framework for Spatio-temporal Analysis and Visualization of Laboratory Mice Tracking Data. Transactions in GIS **11**(5): 765-782.

Kwan, M. P. (2000). Interactive geovisualization of activity-travel patterns using three dimensional geographical information systems: a methodological exploration with a large data set. Transportation Research Part C **8**: 185–203.

Laube, P., et al. (2007). Movement beyond the snapshot – Dynamic analysis of geospatial lifelines. Computers, Environment and Urban Systems **31**: 481 - 501.

Laube, P. and S. Imfeld (2002). Analyzing Relative Motion within Groups of Trackable Moving Point Objects. Lecture notes in computer science, Springer Berlin**:** 132-144.

Laube, P., S. Imfeld and R. Weibel (2005). Discovering relative motion patterns in groups of moving point objects. International Journal of Geographical Information Science **19**(6): 639-668.

Laube, P. and R. S. Purves (2006). An approach to evaluating motion pattern detection techniques in spatio-temporal data. Computers, Environment and Urban Systems **30**(3): 347-374.

Laube, P., M. van Kreveld and S. Imfeld (2004). Finding REMO: detecting relative motion patterns in geospatial lifelines. Proceedings of the 11th International Symposium on Spatial Data Handling, Berlin Heidelberg, DE, Springer: 201-214.

Lee, J. G., et al. (2008). TraClass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering. Proc. 2008 Int. Conf. on Very Large Data Base (VLDB'08), Auckland, New Zealand: 1081-1094.

Lee, J. G., J. Han and K. Y. Whang (2007). Trajectory clustering: a partition-and-group framework. Proceedings of the 2007 ACM SIGMOD international conference on Management of data: 593-604.

Li, X., et al. (2007). Traffic Density-Based Discovery of Hot Routes in Road Networks. Lecture notes in computer science **4605**: 441.

Li, Y., J. Han and J. Yang (2004). Clustering moving objects. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA, ACM New York, NY, USA: 617-622.

Mamoulis, N., et al. (2004). Mining, indexing, and querying historical spatiotemporal data. Proceedings of the 10th ACM International Conference On Knowledge Discovery and Data Mining, ACM: 236-245.

Mark, D. M. and M. J. Egenhofer (1998). Geospatial lifelines. Integrating Spatial and Temporal Database, Dagstuhl Seminar Report No. 228. G. O, S. T and T. B.

Meade, J., D. Biro and T. Guilford (2005). Homing pigeons develop local route stereotypy. Proceedings of the Royal Society B: Biological Sciences **272**(1558): 17.

Miller, H. J. and J. Han (2001). Geographic Data Mining and Knowledge Discovery, CRC Press.

Moore, A. B., et al. (2003). A time geography approach to the visualisation of sport. Proceedings of the 7th International Conference on GeoComputation University of Southampton, United Kingdom.

Morzy, M. (2007). Mining Frequent Trajectories of Moving Objects for Location Prediction. Proceedings of the 5th international conference on Machine Learning and Data Mining in Pattern Recognition. Leipzig, Germany, Springer Berlin**:** 667-680.

Nanni, M. (2002). Clustering Methods for Spatio-Temporal Data. Computer Science Department, University of Pisa. **Ph.D. Thesis**.

Nanni, M., et al. (2008). Spatiotemporal data mining. Mobility, Data Mining and Privacy. Berlin Heidelberg, Springer**:** 267-296.

Nanni, M. and D. Pedreschi (2006). Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems **27**(3): 267-289.

Ng, R. T. (2001). Detecting outliers from large datasets. Geographic data mining and knowledge discovery. H. J. Miller and J. Han**:** 218-235.

Palma, A. T., et al. (2008). A clustering-based approach for discovering interesting places in trajectories. Proceedings of the 2008 ACM symposium on Applied computing. Fortaleza, Ceara, Brazil, ACM**:** 863-868.

Parkes, D. and N. Thrift (1980). Times, Spaces, and Places: A Chronogeographic Perspective. Annals of the Association of American Geographers **71**(2): 292-295.

Pei, J., et al. (2001). PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern. 17th International Conference on Data Engineering (ICDE'01): 215-226.

Perng, C. S., et al. (2000). Landmarks: A New Model for Similarity-Based Pattern Querying in Time Series Databases. 16th International Conference on Data Engineering IEEE Computer Society Press: 33-44.

Pfoser, D. and C. S. Jensen (1999). Capturing the Uncertainty of Moving-Object Representations. Lecture notes in computer science: 111-131.

Porikli, F. (2004). Trajectory distance metric using Hidden Markov Model based representation. IEEE European Conference on Computer Vision, PETS Workshop.

Preisler, H. K., et al. (2004). Modeling animal movements using stochastic differential equations Environmetrics **15**: 643-657.

Qi, F. and A. X. Zhu (2003). Knowledge discovery from soil maps using inductive learning. International Journal of Geographical Information Science **17**(8): 771-795.

Raper, J. (2002). The dimensions of GIScience. 2ed International Conference of Geographic Information Science, GIScience 2002.

RMEF. (1999). Rocky Mountain Elk Foundation. Retrieved December 15, 2008, from http://www.rmef.org/home.

Roberts, S., et al. (2004). Positional entropy during pigeon homing I: application of Bayesian latent state modelling. Journal of Theoretical Biology **227**(1): 39-50.

Roddick, J. F. and B. G. Lees (2001). Paradigms for spatial and spatio-temporal data mining. Geographic data mining and knowledge discovery. H. J. Miller and J. Han, Taylor & Francis**:** 33-50.

Rowland, M. M., et al. (1997). The Starkey project: history, facilities, and data collection methods for ungulate research. U.S. Forest Service General Technical Report PNW-GTR-396. Portland, OR, US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Rowland, M. M., et al. (1998). The Starkey habitat database for ungulate research: construction, documentation, and use. U.S. Forest Service General Technical Report PNW-GTR-430. Portland, OR, US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Shekhar, S. and S. Chawla (2003). Spatial Databases: A Tour, Prentice Hall.

Shim, C. B. and J. W. Chang (2003). Efficient similar trajectory-based retrieval for moving objects in video databases. Image and Video Retrieval CIVR, Springer Berlin. **2728:** 613-618.

Shoshany, M., A. Even-Paz and S. Bekhor (2007). Evolution of clusters in dynamic point patterns: with a case study of Ants' simulation. International Journal of Geographical Information Science **21**(7): 777–797.

Song, L., et al. (2006). Evaluating Next-Cell Predictors with Extensive Wi-Fi Mobility Data. IEEE Transactions on Mobile Computing **5**(12): 1633-1649.

Spaccapietra, S., et al. (2008). A conceptual view on trajectories. Data & Knowledge Engineering **65**: 126-146.

Tversky, B., J. B. Morrison and M. Betrancourt (2002). Animation: can it facilitate? International Journal of Human Computer Studies **57**(4): 247-262.

Ultsch, A. and H. P. Siemon (1990). Kohonen's self organizing feature maps for exploratory data analysis. Proceedings of the International Neural Network Conference, Dordrecht, Netherlands, Kluwer: 305-308.

US Forest Service. (1996). The Starkey Project. Retrieved October 8, 2008, from http://www.fs.fed.us/pnw/starkey/data/tables/index.shtml.

Verhein, F. and S. Chawla (2006). Mining Spatio-temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases. Lecture Notes in Computer Science(3882): 187-201.

Vlachos, M., D. Gunopulos and G. Das (2004). Rotation invariant distance measures for trajectories. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA, ACM New York, NY, USA: 707-712.

Vlachos, M., et al. (2003). Indexing multi-dimensional time-series with support for multiple distance measures. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data minin, Washington, D.C., ACM: 216-225.

Vlachos, M., G. Kollios and D. Gunopulos (2002). Discovering Similar Multidimensional Trajectories. 18th International Conference on Data Engineering (ICDE'02), IEEE Computer Society Press; 1998: 673-684.

Weimerskirch, H., et al. (2002). GPS Tracking of Foraging Albatrosses. Science **295**(5558): 1259-1259.

Wilensky, U. (1998). NetLogo Ants model, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. Retrieved October 6, 2008, from http://ccl.northwestern.edu/netlogo/models/Ants.

Wolfer, D. P., et al. (2001). Extended analysis of path data from mutant mice using the public domain software Wintrack. Physiology and Behavior **73**: 745–753.